Hierarchical modeling

David W. Hogg
Center for Cosmology and Particle Physics, New York University

2011 April 2

Polemic: Sometimes it is the *prior* that we seek

- We know about thousands of exoplanets, each of which has a period T.
- Do we care about any particular planet's period?
  - Yes, sometimes: We might want to schedule observations, or estimate habitability.
  - No, usually: We want to understand the processes that generate the *distribution* of periods.
- We want to know the true distribution from which periods are drawn.
- This true distribution is what we should be using as the *prior* in every individual planet inference.
- Can we parameterize and infer a *prior*?

Conclusions

- Hierarchical modeling is simple, powerful, and generic.
  - Some of you are using it already (some without knowing it).
- We have obtained powerful results with it.
  - eccentricity distributions for exoplanets
  - classification: quasar target selection
  - prediction: photometric redshifts
- It is a form of *deconvolution* and we shouldn't be afraid of that.

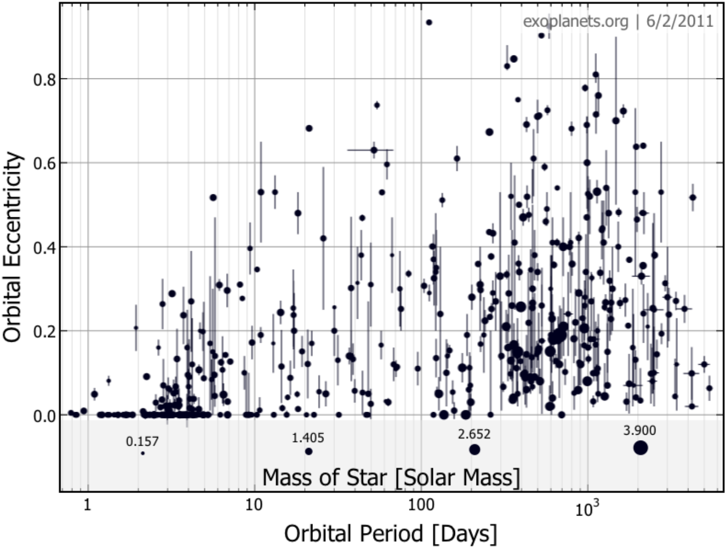
Principal collaborators

- **Jo Bovy** (NYU $\rightarrow$ IAS)
- Joe Hennawi (MPIA)
- **Dustin Lang** (Princeton)
- Adam Myers (UIUC $\rightarrow$ Wyoming)
- Hans-Walter Rix (MPIA)
- Sam Roweis (deceased)
- *SDSS-III* Collaboration

Eccentricity estimation

- Single-point (*e.g.*, maximum-likelihood) eccentricity estimates are biased high.
  - Shen & Turner (2008); others
  - comes from model freedom: higher $e \rightarrow$ greater model freedom
  - (recall continous model complexity)
- Most MCMC or Bayesian approaches use *demonstrably wrong* flag priors on e.
- What priors should we be using?
  - even if we use a justified prior, single-point estimates will always be bad
- It matters!

Eccentricities

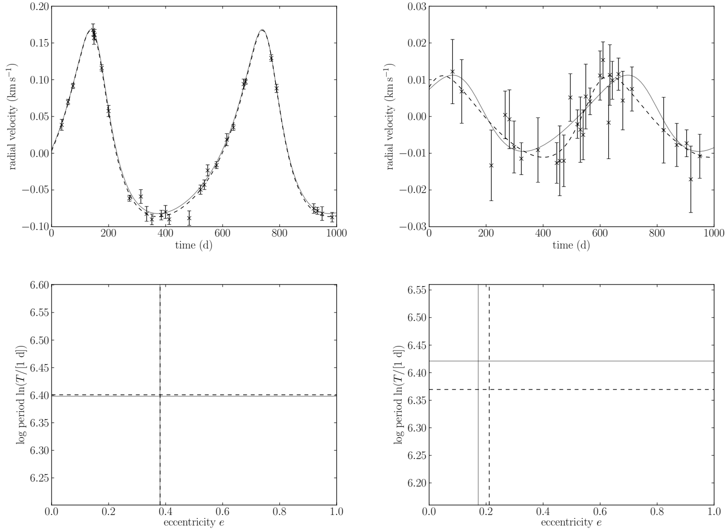


Eccentricity inference, usual story

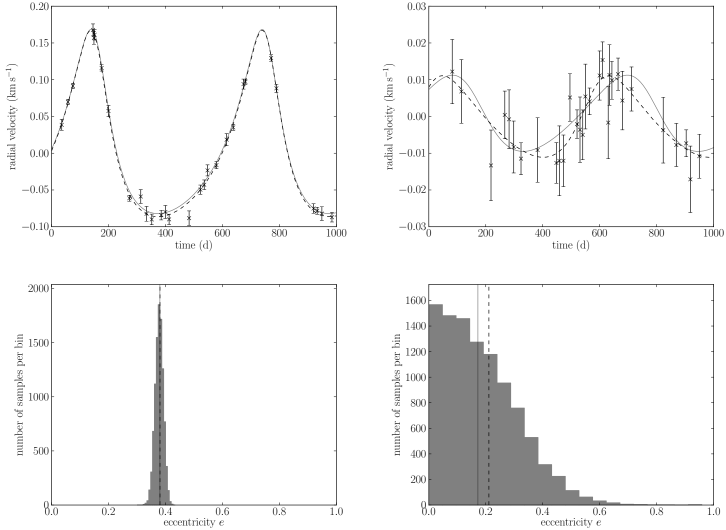
$$
\begin{aligned}
\boldsymbol{\omega}_n &\equiv (\kappa_n, T_n, \phi_n, e_n, \varpi_n) \\
v_{nj} &= V_n + g_{\boldsymbol{\omega}_n}(t_{nj}) + E_{nj} \\
-2\,\ln p(\mathbf{D}_n|\boldsymbol{\omega}_n) &= Q + \sum_{j=1}^{M_n} \ln(\sigma_{nj}^2 + S_n^2) + \sum_{j=1}^{M_n} \frac{[V_n + g_{\boldsymbol{\omega}_n}(t_{nj}) - v_{nj}]^2}{\sigma_{nj}^2 + S_n^2} \\
p(\boldsymbol{\omega}_n|\mathbf{D}_n) &= \frac{1}{Z_n}\, p(\mathbf{D}_n|\boldsymbol{\omega}_n)\, p_0(\boldsymbol{\omega}_n) \quad ,
\end{aligned}
$$

where $p_0(\boldsymbol{\omega}_n)$ is some "uninformative" prior like flat in some parameters, $1/x$ in others.

Eccentricity inference demo



Eccentricity inference demo



Eccentricity distribution inference (1008.4146)

What if you think there might be some family of priors $p(\boldsymbol{\omega}_n|\boldsymbol{\alpha})$ parameterized by some $\boldsymbol{\alpha}$; could you infer this?

$$
\begin{aligned}
p(\{\mathbf{D}_n\}_{n=1}^N \,|\, \{\boldsymbol{\omega}_n\}_{n=1}^N) &= \prod_{n=1}^N p(\mathbf{D}_n|\boldsymbol{\omega}_n) \\
p(\{\mathbf{D}_n\}_{n=1}^N \,|\boldsymbol{\alpha}) &= \prod_{n=1}^N \int \mathrm{d}\boldsymbol{\omega}_n \, p(\mathbf{D}_n|\boldsymbol{\omega}_n)\, p(\boldsymbol{\omega}_n|\boldsymbol{\alpha}) \quad .
\end{aligned}
$$

This is still a likelihood, but we have marginalized out the properties of every exoplanet—these are “nuisance” parameters in this formulation.

Eccentricity distribution inference (1008.4146)

Say all you get, for each exoplanet, are K samples drawn from an uninformative prior. What then? Importance sampling.

$$
\begin{aligned}
p(\boldsymbol{\omega}_n|\boldsymbol{\alpha}) &\equiv \frac{f_{\boldsymbol{\alpha}}(e_n)\,p_0(\boldsymbol{\omega}_n)}{p_0(e_n)} \\
\int \mathrm{d}\boldsymbol{\omega}_n\,p_0(\boldsymbol{\omega}_n|\mathbf{D}_n)\,F(\boldsymbol{\omega}_n) &\approx \frac{1}{K}\sum_{k=1}^{K} F(\boldsymbol{\omega}_{nk}) \\
p(\{\mathbf{D}_n\}_{n=1}^{N}\,|\boldsymbol{\alpha}) &\approx \prod_{n=1}^{N}\frac{1}{K}\sum_{k=1}^{K}\frac{f_{\boldsymbol{\alpha}}(e_{nk})}{p_0(e_{nk})}
\end{aligned}
$$

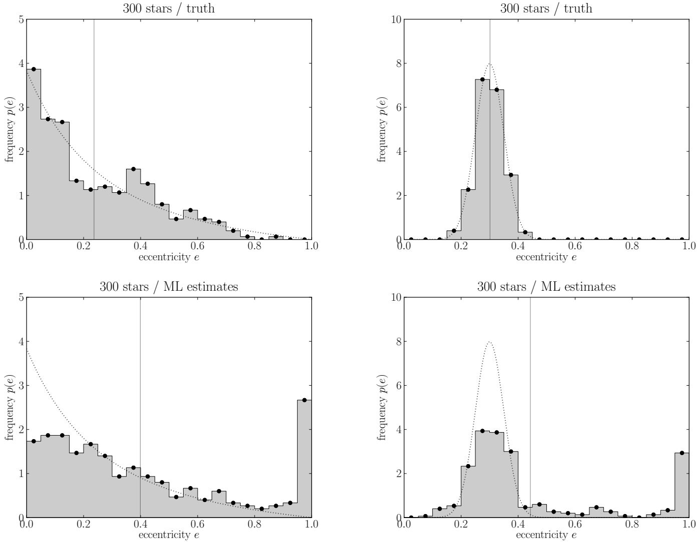
Eccentricity distribution model (1008.4146)

Use a non-parametric (read: very highly parameterized) function for the eccentricity distribution): Step function with M steps.

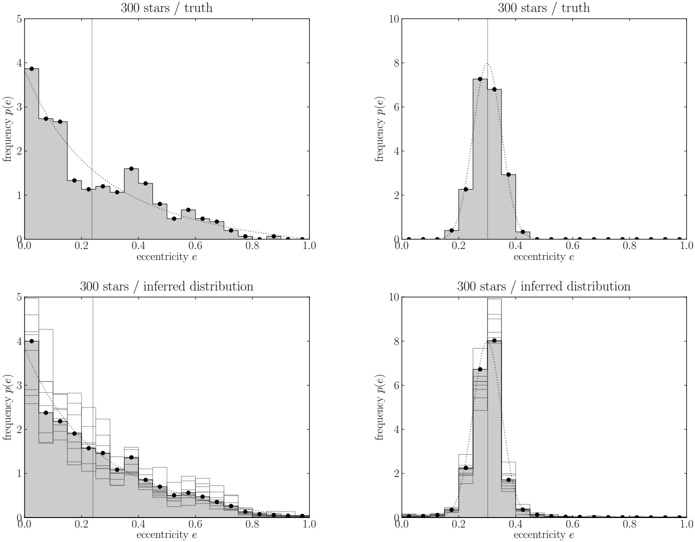
$$
\begin{aligned}
f_{\boldsymbol{\alpha}}(e) &\equiv \sum_{m=1}^{M} \exp(\alpha_m)\, s(e; \tfrac{m-1}{M}, \tfrac{m}{M}) \\
s(x; L, H) &\equiv \begin{cases} 0 & \text{for } x < L \\ (H-L)^{-1} & \text{for } L \leq x \leq H \\ 0 & \text{for } H < x \end{cases} \\
\sum_{m=1}^{M} \exp \alpha_m &= 1 \\
p(\boldsymbol{\alpha}) &\propto \delta(1 - \sum_{m=1}^{M} \exp \alpha_m)\, \exp(-\frac{1}{2}\, \epsilon \sum_{m=2}^{M} [\alpha_m - \alpha_{m-1}]^2)
\end{aligned}
$$

Note Gaussian-processes-like regularization.

Distribution inference demo: ML estimates—bad



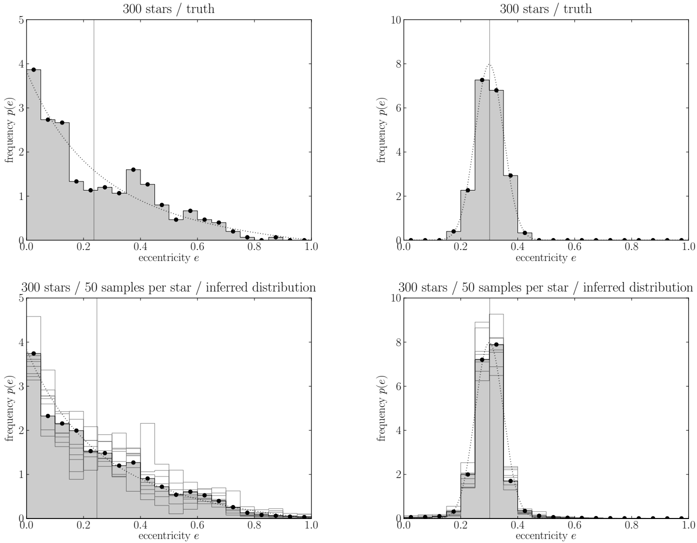
Distribution inference demo: Good!



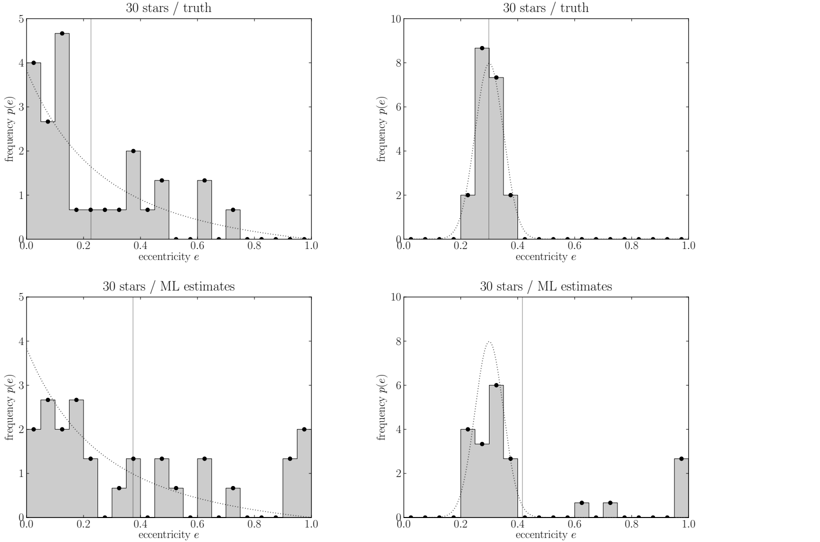
Polemic: Deconvolution

- We can infer the true distribution even with extremely noisy measurements.
- This is an extreme form of *deconvolution*.
  - (but not *Extreme Deconvolution (tm)*)
- Depends crucially on having full—and accurate—likelihood or posterior information.
- Performed by "forward modeling".

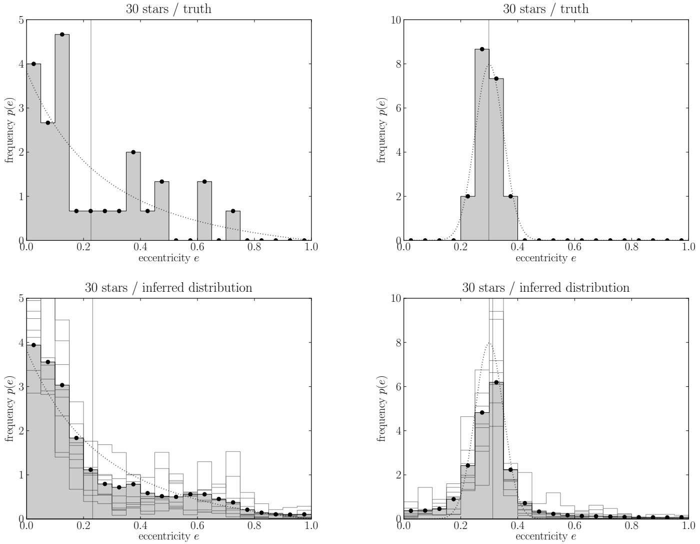
Distribution inference demo: Small samplings



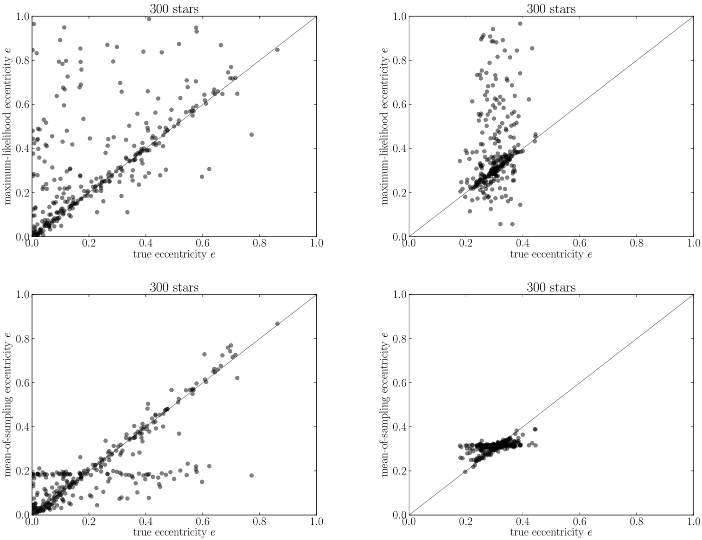
Distribution inference demo: Small sample



Distribution inference demo: Still good!



Distribution inference demo: Truly hierarchical



Conclusions

- Hierarchical modeling is simple, powerful, and generic.
  - Some of you are using it already (some without knowing it).
- We have obtained powerful results with it.
  - eccentricity distributions for exoplanets
  - classification: quasar target selection
  - prediction: photometric redshifts
- It is a form of *deconvolution* and we shouldn't be afraid of that.

Quasar target selection: setup

- $2.2 < z < 3.5$ quasars can be used to measure the baryon acoustic oscillation in the Lyman alpha forest
- *SDSS-III BOSS*
- quasars in this range *look like stars* in *ugriz*
- This is a hard supervised classification problem.

What's wrong with typical classification algorithms?

- neural networks, boltzmann machines, support vector machines, boosting
- these are all *awesome*
- they require that *test data* have the same statistical and error properties as *training data*

- they require that all features be measured for all data points

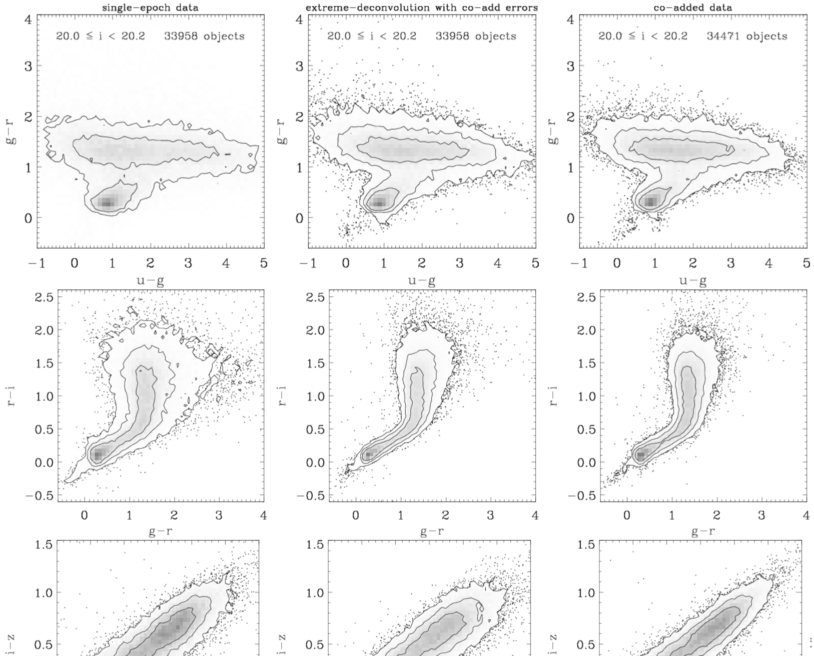
What's wrong with typical classification algorithms?

- neural networks, boltzmann machines, support vector machines, boosting
- these are all *awesome*
- they require that *test data* have the same statistical and error properties as *training data*
- *never true!*
- they require that all features be measured for all data points
- *never true!*

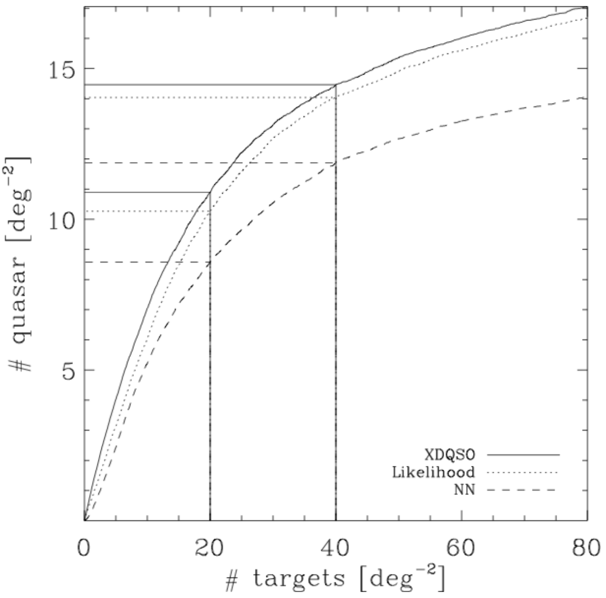
XDQSO target selection (1011.6392): Method

- extreme deconvolution:
- each data point samples the true density (in color space), *convolved* with that data point's own unique uncertainty profile
  - an independent and unique convolution of the model for every data point
  - like having as many classifiers as data points
- model all this with mixtures of Gaussians for performance
- likelihood ratios (star *vs.* galaxy) are density ratios in the convolved model

XDQSO target selection (1011.6392): Results



XDQSO target selection (1011.6392): Results



XDQSO target selection (1011.6392): why we are so good?

- We use the errors correctly and account properly for missing data; we have a *generative model*.
- That is true for both the training data and the test data.
- We are extensible to new prior information or other data.
  - *GALEX*
  - *UKIDSS*
  - variability
- Bovy
- *extreme-deconvolution* (at code.google.com)
  - Bovy, Hogg, & Roweis (0905.2979)
  - it Just Works (tm)
  - C code with Python and IDL wrappers / interface
  - can handle large data sets with large numbers of dimensions
- *SDSS-III BOSS* core target selection

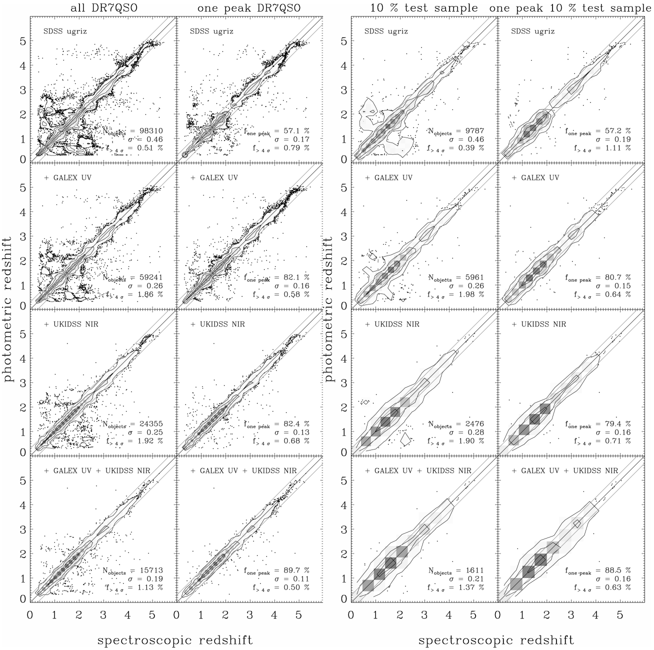
Polemic: Missing data

- Most machine-learning methods hate missing data.
- Interpolation or data censoring (both very, very bad) are required.
- Any model that properly accounts for *uncertainty* also properly accounts for *missing data*.
    - Missing data is (extreme) uncertainty; uncertainty is (mild) missing data.
- If you have a justified generative model $p(\mathbf{D}_n|\omega_n)$, you automatically deal with missing data.

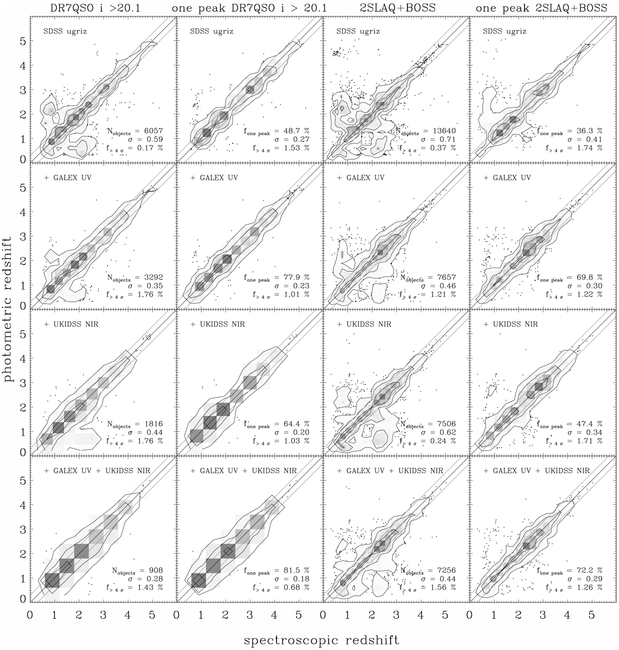
XDQSOz redshift prediction (1105.3975)

- Add redshift as a dimension to the photometric XDQSO.
- Add also *GALEX* and *UKIDSS*.
  - Not full coverage? No problem!
- Model with *extreme deconvolution* again.
- Condition model on available photometry and predict redshift.
  - Not all bands measured? No problem!

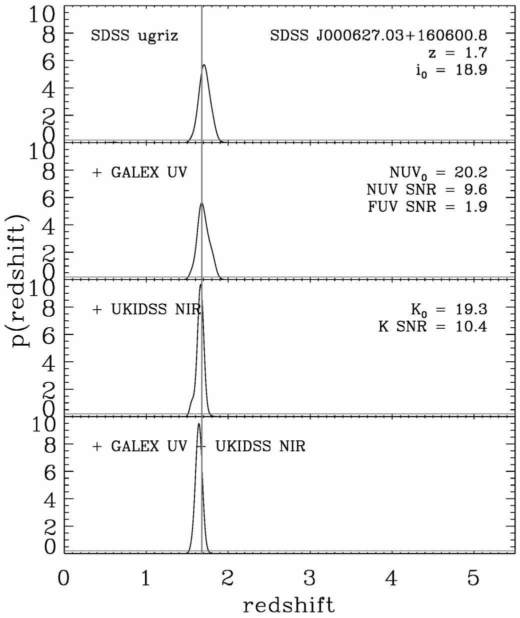
XDQSOz redshift prediction (1105.3975): Results



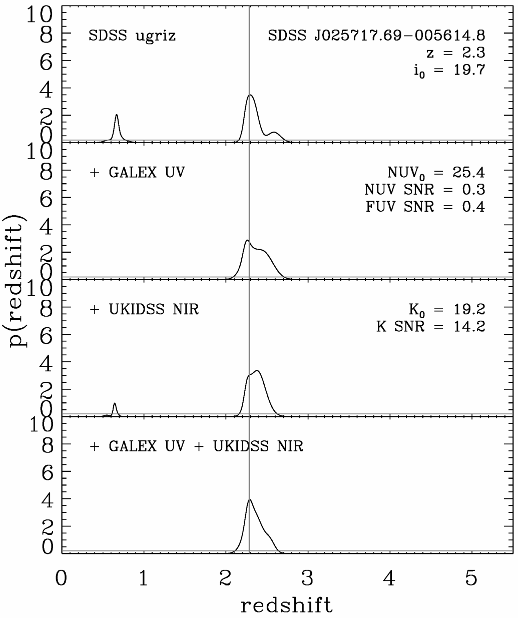
XDQSOz redshift prediction (1105.3975): Results



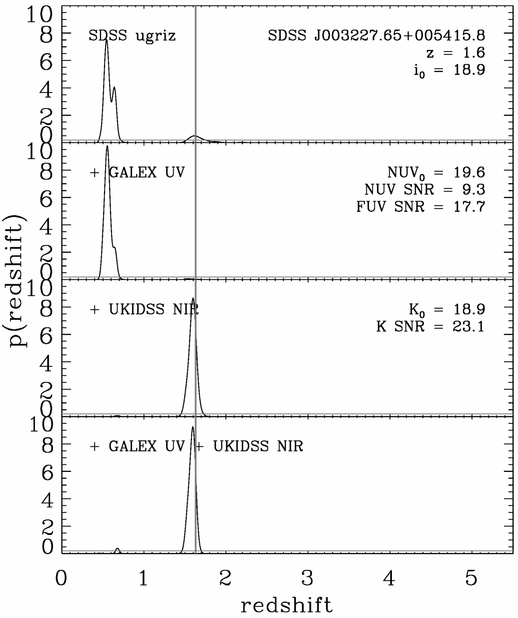
XDQSOz redshift prediction (1105.3975): Examples



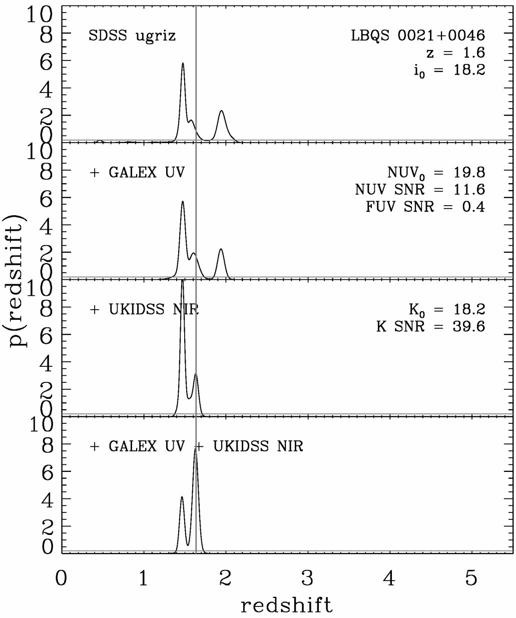
XDQSOz redshift prediction (1105.3975): Examples



XDQSOz redshift prediction (1105.3975): Examples



XDQSOz redshift prediction (1105.3975): Examples



XDQSOz redshift prediction (1105.3975): Results

- We have the most precise and accurate photometric redshift estimates for quasars in the magnitude and redshift ranges relevant to *SDSS-III BOSS*.
- We can use all photometric bands where they are available, but don't need complete data.
- Signal-to-noise of training and test sets do not have to be similar.
- Makes great use of extremely low signal-to-noise *GALEX* data in both training and testing.

Polemic: Don't convolve your data, convolve your model!

- If you are uncertain about something (a redshift, a classification) so that you don't know which bin to put it in:
- *don't* put a bit of it into each bin!
  - That re-convolves your noisy result with the noise again.
- *Do* put a bit of your *distribution model* into each bin.
  - That is, convolve your *model* for the object with the uncertainty.
  - Obvious, but easy to get wrong.

Conclusions

- Hierarchical modeling is simple, powerful, and generic.
  - Some of you are using it already (some without knowing it).
- We have obtained powerful results with it.
  - eccentricity distributions for exoplanets
  - classification: quasar target selection
  - prediction: photometric redshifts
- It is a form of *deconvolution* and we shouldn't be afraid of that.