How I expect to access the *Gaia* Catalog

David W. Hogg

Center for Cosmology and Particle Physics, New York University

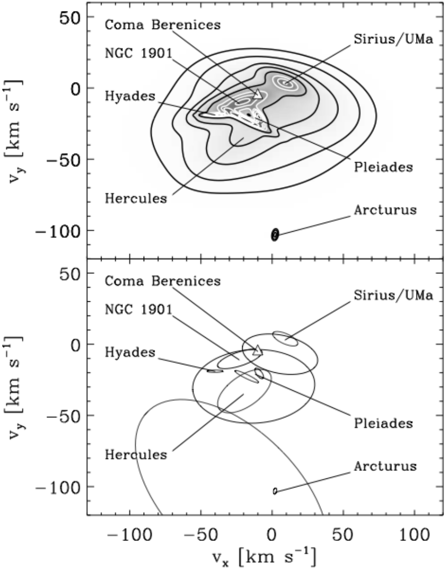
2010 June 11

summary

- (sensibly) I propose a definition of catalog-entry *uncertainty*.
- (radically) I recommend a *sampling* of *Gaia* Catalogs.
- (insanely) I recommend [redacted]
- These suggestions are motivated by scientific considerations.

Cosmology with *Gaia*

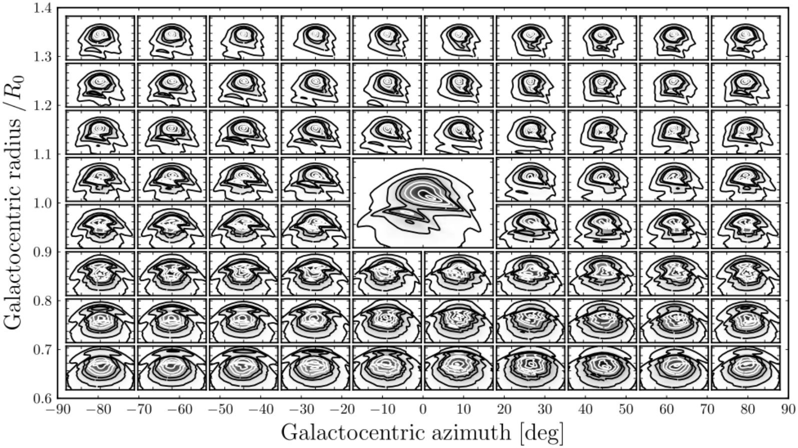
- The Milky Way is the best place to study dark matter at small scales and in the nonlinear regime.
- If dark matter annihilation is tentatively detected, can we confirm non-trivialities through dynamical tests?
- We expect *extremely rich structure* in the Galaxy's dark sector; if there is no annihilation signal, dynamics is our *only tool*.
  - think: informative, coherent phase-space structure
  - think: dynamical memory of encounters and perturbations
- *Precise* experiments must be done probabilistically (that is, with likelihoods or worse).



Bovy, Hogg, & Roweis 2009 *ApJ* **700** 1794–1819

Extreme deconvolution

- Estimating a distribution function given noisy observations?
- Every data point has its own special error properties.
- Every data point can be missing some dimensions.
- Want the distribution function that maximizes the probability of each data point, when convolved with each data point's unique uncertainty properties.
  - note frequentism?
- The best possible method, but it needs good uncertainty estimates.
  - Bovy, Hogg, & Roweis, arXiv:0905.2979
  - http://code.google.com/p/extreme-deconvolution/
- Can set model complexity by cross-validation.
  - note frequentism?



Bovy 2010 arXiv:1006.0736

Stream-finding

- In phase space, clusters evolve to (finite-thickness) one-dimensional lines.
  - three conserved quantities—actions
  - two unexplored directions in angle space
- *Gaia* will find thousands of these, by any estimate.
- The faintest require a multi-star hypothesis test for validation.
  - A small covariance (in, say, the radial velocities) can *dominate* this hypothesis test once there are many stars being tested simultaneously.

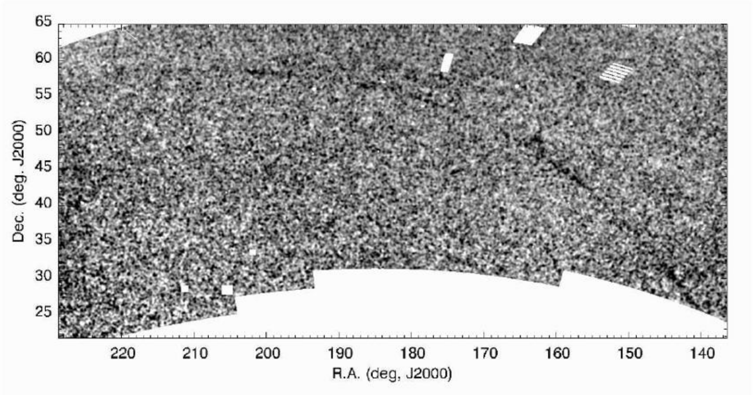
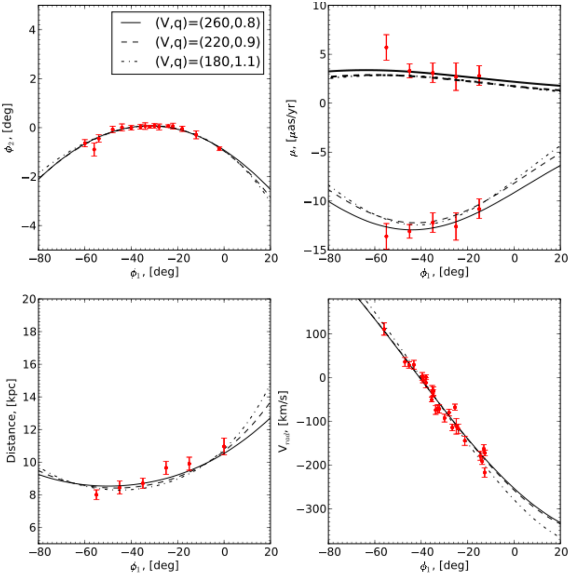


Fig. 1.— Smoothed, summed weight image of the SDSS field after subtraction of a low-order polynomial surface fit. Darker areas indicate higher surface densities. The weight image has been smoothed with a Gaussian kernel with $\sigma = 0.2°$. The white areas are either missing data, or clusters, or bright stars which have been masked out prior to analysis.

Grillmair & Dionatos 2006 *ApJL* **643** L17–L20.



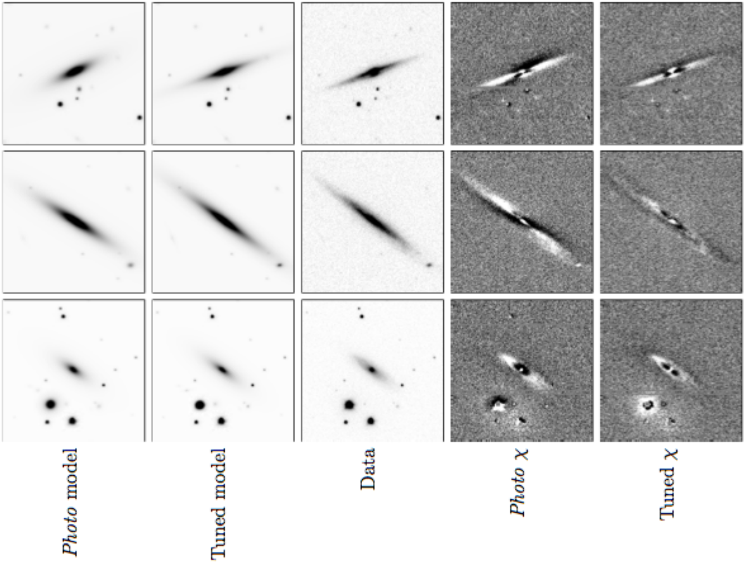
Koposov, Rix, & Hogg, 2010 *ApJ* **712** 260–273.

Full Milky-Way modeling

- Generate models of the observed distribution of stars given a dynamical model and a distribution function.
- Comparison of models is by necessity a full-Catalog (or nearly so) multi-star hypothesis test.

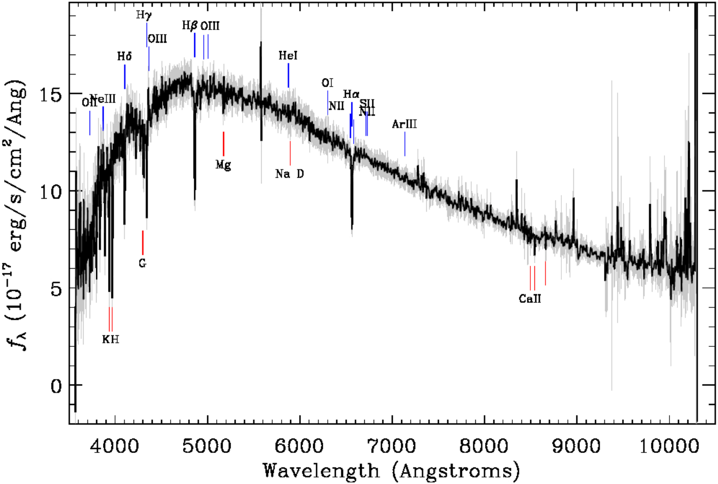
Polemic: Telescopes do *not* generate *catalogs*

- . . . they generate *intensity measurements!*
- We want to test models against the intensity measurements.
- A well-designed catalog permits this.
- Hogg & Lang, *The theory of everything*, arXiv:0810.3851



Lang, Hogg & Peng, *NIPS* submitted

Survey: *boss* Program: *boss* Target: *STD_FSTAR*
RA=209.83191, Dec=1.08057, Plate=4036, Fiber=84, MJD=55330
cz=−60+/−5 km/s Class=STAR F5
No warnings.



Spectro-perfectionism

- an *SDSS-III* spectrum: $\lambda_i, f_i, 1/\sigma_i^2$
- These are *not* (just) measurements of flux with standard errors!
- two models: $m_1(\lambda)$, $m_2(\lambda)$
- define

$$\Delta\chi^2 \equiv \sum_i \left[\frac{m_2(\lambda_i) - f_i}{\sigma_i^2}\right] - \sum_i \left[\frac{m_1(\lambda_i) - f_i}{\sigma_i^2}\right]$$

- *Define* $f_i, 1/\sigma_i^2$ so that this is *as close as possible* to what you would have computed for $\Delta\chi^2$ in the *read-out spectrograph image pixels*.
  - Bolton & Schlegel, arXiv:0911.2689
- outputs are *parameters* of a Gaussian approximation to the likelihood function!

Sensible proposal: Uncertainty definition

- a *Gaia* catalog entry: $\mathbf{y}^{\mathsf{T}} = [\mathrm{RA}, \mathrm{Dec}, \pi, \mu_\alpha, \mu_\delta, v_r]$, $\mathbf{C}^{-1}$
- Two hypotheses: $\mathbf{Y}_1$, $\mathbf{Y}_2$
- define

$$\Delta\chi^2 \equiv [\mathbf{Y}_2 - \mathbf{y}]^{\mathsf{T}} \cdot \mathbf{C}^{-1} \cdot [\mathbf{Y}_2 - \mathbf{y}] - [\mathbf{Y}_1 - \mathbf{y}]^{\mathsf{T}} \cdot \mathbf{C}^{-1} \cdot [\mathbf{Y}_1 - \mathbf{y}]$$

- *Define* $\mathbf{y}, \mathbf{C}^{-1}$ so that this is *as close as possible* to what you would have computed for $\Delta\chi^2$ in the *telemetered image pixels*, marginalizing over all nuisance parameters.
  - a marginalized likelihood?
  - see hierarchical Bayes literature
  - Gelman *et al.*, *Bayesian Data Analysis* (Chapman & Hall)

Polemic: Publish likelihoods, not posteriors!

- Yes, Bayes's rule is the right way to do inference, but:
- Data enter inference through the likelihood.
- Different users have different priors (because they have different data).
- Subsequent users want to combine (say) three datasets without *cubing* the prior.
- Even if you *insist* on publishing posteriors, *also* publish the prior, so subsequent users can divide it out.

Radical proposal: A sampling of *Gaia* Catalogs

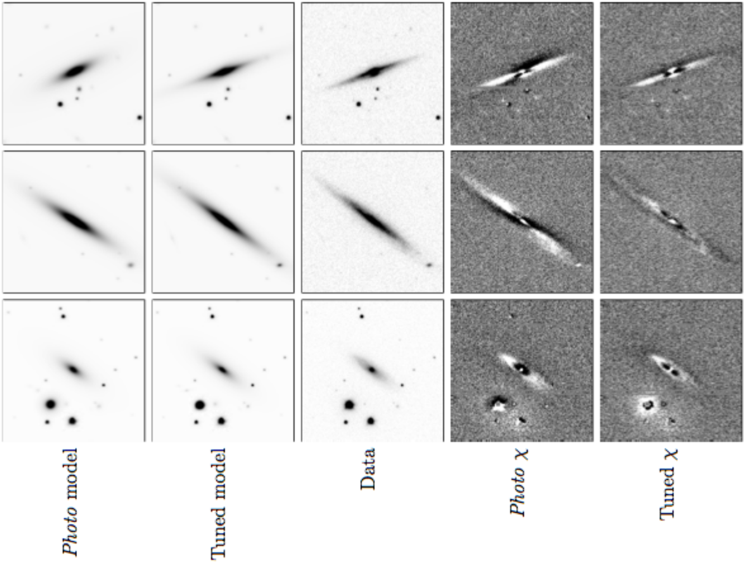
- Make not one *Gaia* Catalog but $K + 1$.
- The “zeroth” Catalog is your principal release.
- The other K are samples from a posterior distribution, with
  - astrometric catalog entry variations, and
  - calibration nuisance parameter variations
- such that an *average* of any quantity *over K samples* is close to a marginalization over *all* probabilistic quantities.
- This is equivalent to a rank-K approximation of the covariance matrix
  - *cf.* Holl contribution

Radical proposal: Permit qualitative changes

- There is *no need* to make *hard decisions* even about qualitatively different models.
  - 0, 1, or N expolanets?
  - binary star?
- return sampling for each possibility, plus likelihoods
- empowers users:
  - different users have different *priors*
  - different users have different *utilities*
  - different users have different *data* (which they want to combine with *Gaia* data optimally)

Insane proposal: Expose the likelihood function

- Input: (catalog, nuisance-parameter) *diff*
- Output: $\Delta \log \mathscr{L}$
  - permits any user to compute *any* element of the full covariance
  - could shift computational burden to users
  - challenging now; easy in 2020
  - annoyed? hey, talk is cheap!



Lang, Hogg & Peng, *NIPS* submitted

summary

- (sensibly) I propose a definition of catalog-entry *uncertainty*.
- (radically) I recommend a *sampling* of *Gaia* Catalogs.
- (insanely) I recommend exposing the likelihood function.
- These suggestions are motivated by scientific considerations.