# Data-driven models of the Milky Way in the Gaia era

Boris Leistedt — @ixkael, www.ixkael.com
CCPP, New York University

# Happy collaborators

David Hogg
(NYU/Flatiron)

Axel Widmark
(Stockholm)

Lauren Anderson
(Flatiron)

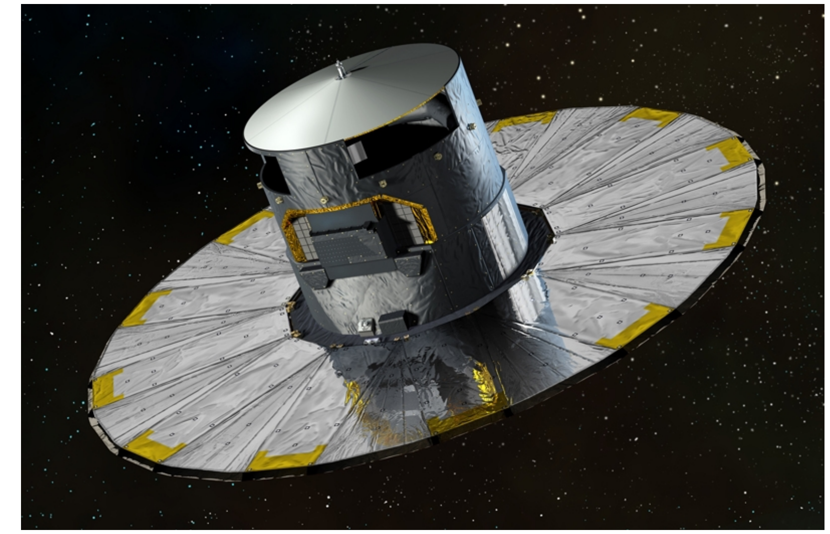
Keith Hawkins
(Columbia)

Adrian Price-Whelan
(Princeton)

Jo Bovy
(Toronto/Flatiron)

# The Gaia mission

www.cosmos.esa.int/web/gaia/science-performance



Successor to Hipparcos

Micro-arcsecond global astrometry for 1+ billion stars, complete to 20th mag: correlated positions, proper motions, parallaxes, apparent mags (3 broad photometric bands).

Radial velocities (NIR medium-res $\lambda/\Delta\lambda$=11k integral-field spectrograph) down to $G_{RVS} \approx 16$ mag

Powerful synergies with other surveys (2MASS, WISE, SDSS, etc)

Many science goals! Solar, Galactic, and extra-Galactic.

# Gaia sprints

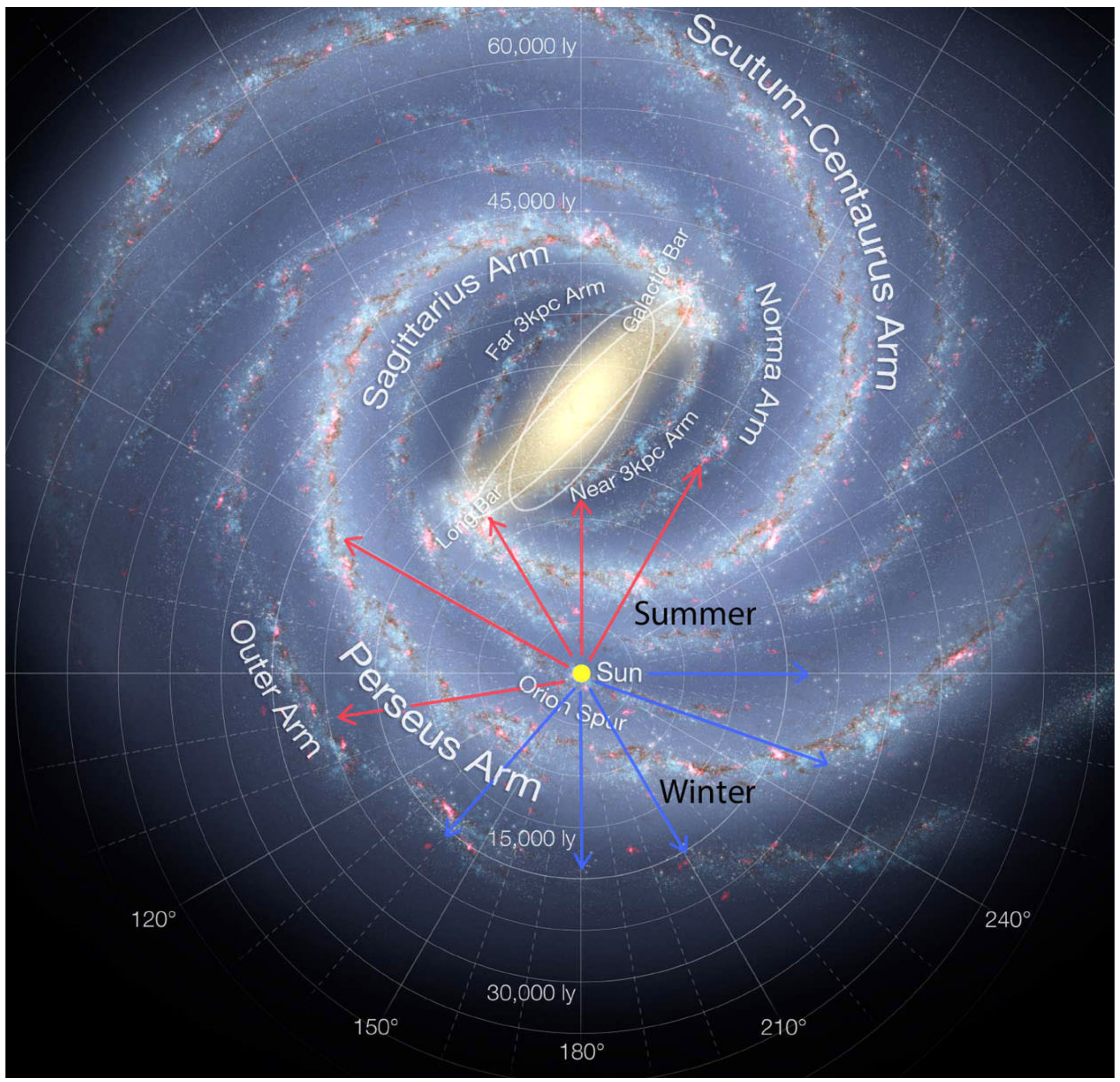
http://gaia.lol

Full week of sprinting/hacking
on concrete achievable projects,
in a room full of experts.

- October 2016 in NYC
- July 2017 in MPIA Heidelberg
- June 2018 in NYC

Dozens of papers & new collaborations!

# Detailed 3D Milky Way models with Gaia




- stellar density (poisson process) and potential
- dust and total dust extinction
- dynamics: full phase-space
- correlation between phase-space & stellar parameters

# Methodological challenges

Correct and full exploitation of Gaia
= difficult regime for data analysis and inference

- Huge data set where uncertainties matter (e.g., magnitudes, parallaxes, proper motions)

- Constraining power of the data exceeds quality of existing physical models (e.g., 3D density, etc). *Worse: using those models can bias the data analysis.*

# Our goals

- **Correct usage of all of the data** (with uncertainties, correlations, selection effects, etc) for new discoveries.

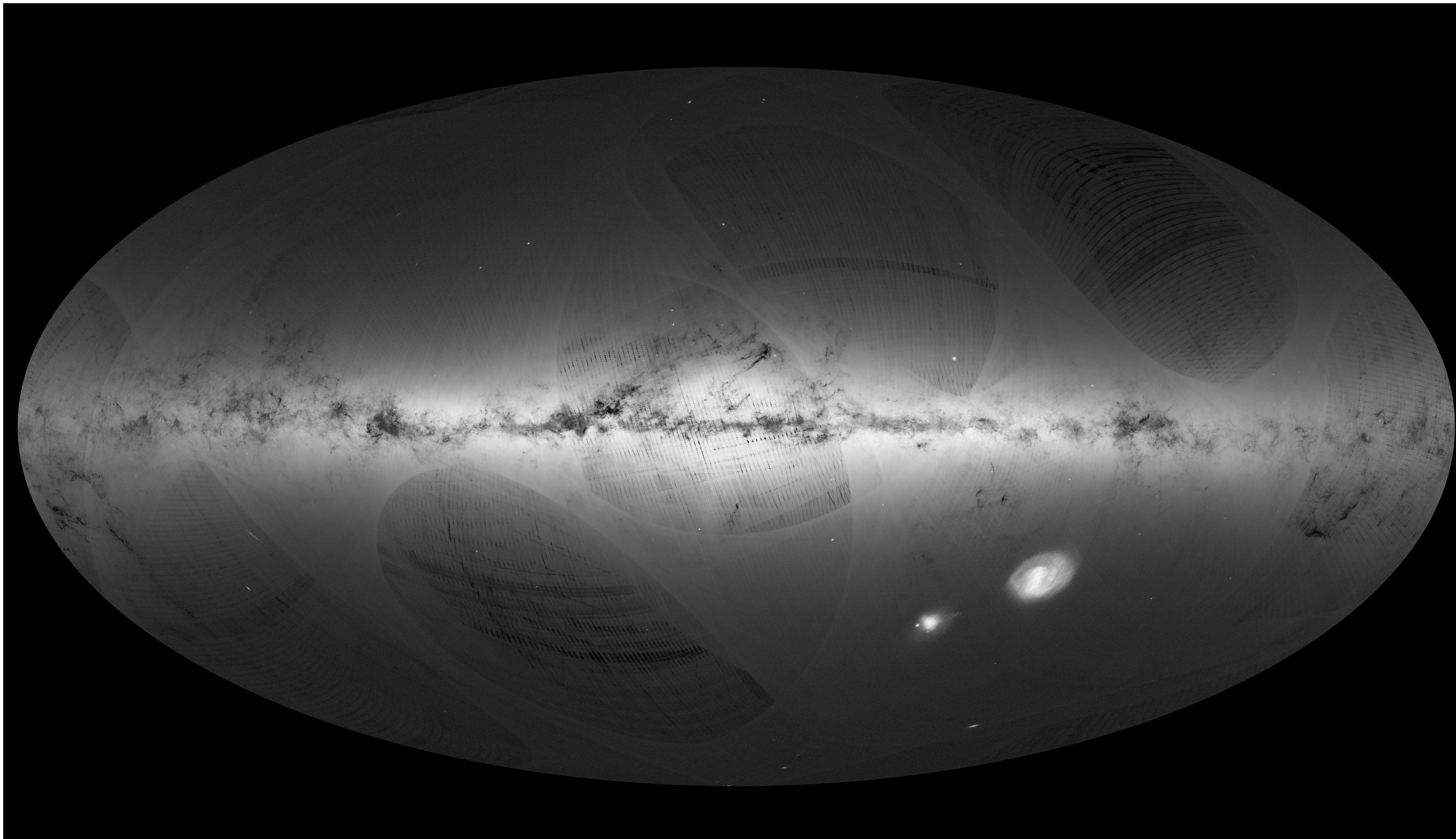
- Develop flexible **"data-driven" models** (e.g., non-parametric) which will inform physical models.

- Gaia DR1 projects:
  - probabilistic models of the color-magnitude diagram
  - calibration of red-clump stars as standard candles
  - improved distance estimates for all Gaia stars
  - detection of unresolved double+triple sequences

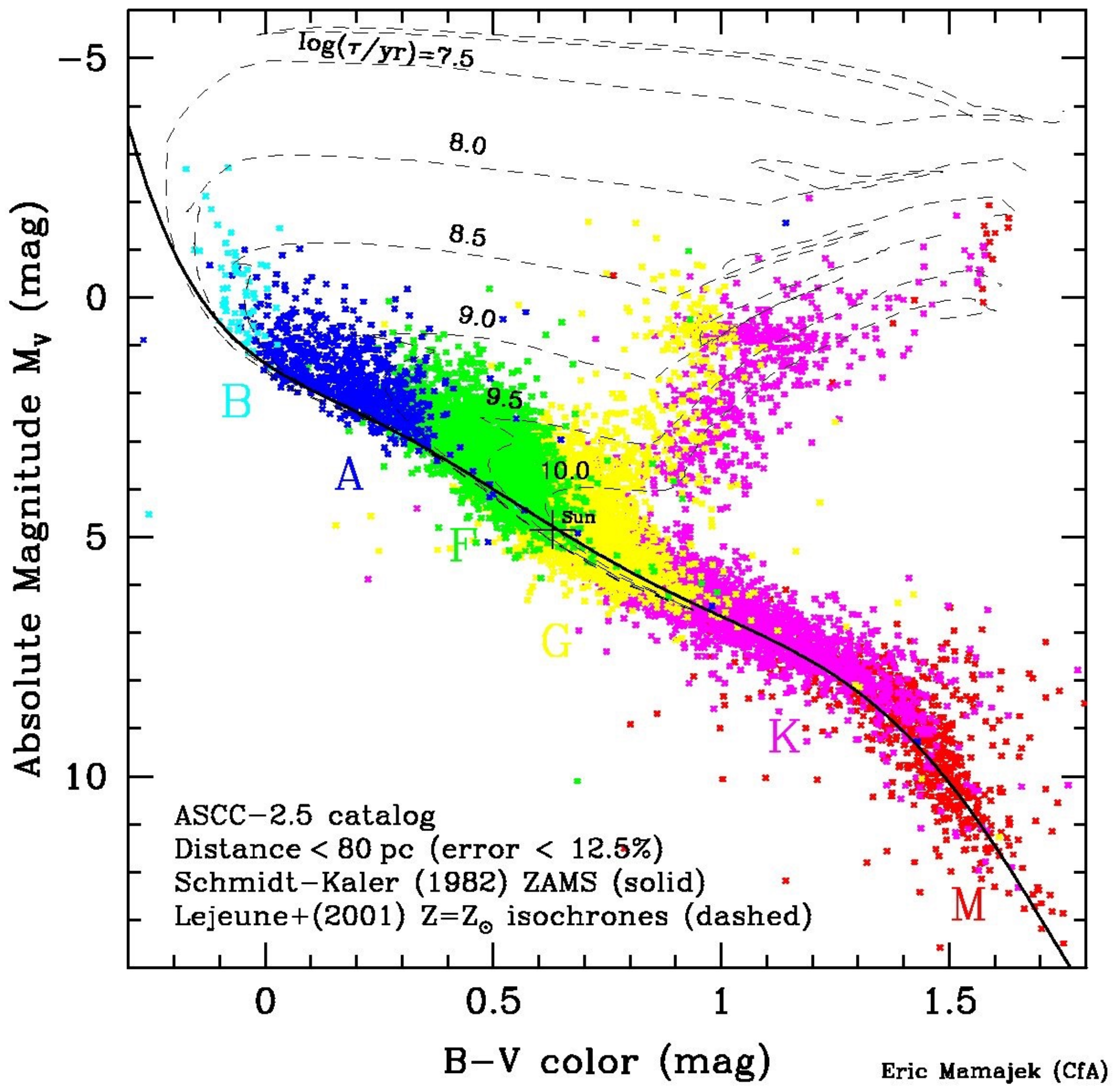
- Gaia DR2: exciting developments, see final slide.

# Gaia Data Release 1

Positions for all sources, but astrometric solution for 2e6 objects in Tycho-2



# There is distance information in magnitudes




Absolute magnitude:

$$M_V = m_V - 5\log_{10}\left(\frac{d}{10\text{ pc}}\right)$$

Parallax & magnitude likelihoods:

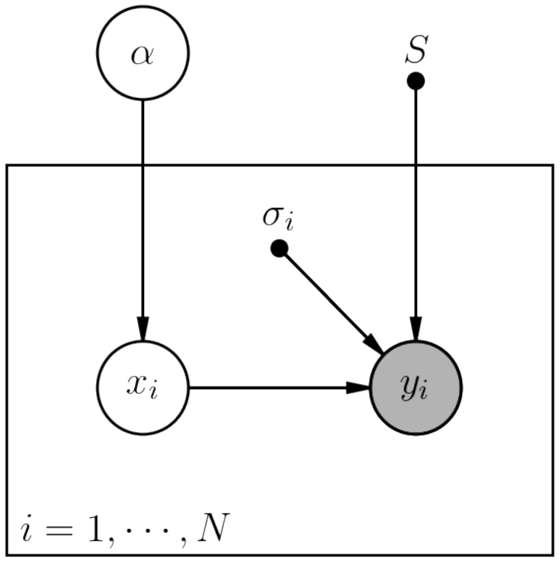
$$p(\hat{\varpi}|d,\sigma_\varpi) = \mathcal{N}(\hat{\varpi} - 1/d; \sigma_\varpi^2)$$

$$p(\hat{\vec{m}}|d,\vec{C},M,\Sigma_{\hat{\vec{m}}}) = \mathcal{N}\left(\hat{\vec{m}} - \vec{m}(d,\vec{C},M); \Sigma_{\hat{\vec{m}}}\right)$$

How to tap into that information without external data/models?

*Construct a color-magnitude diagram from the Gaia data alone!*

# Hierarchical probabilistic models 101



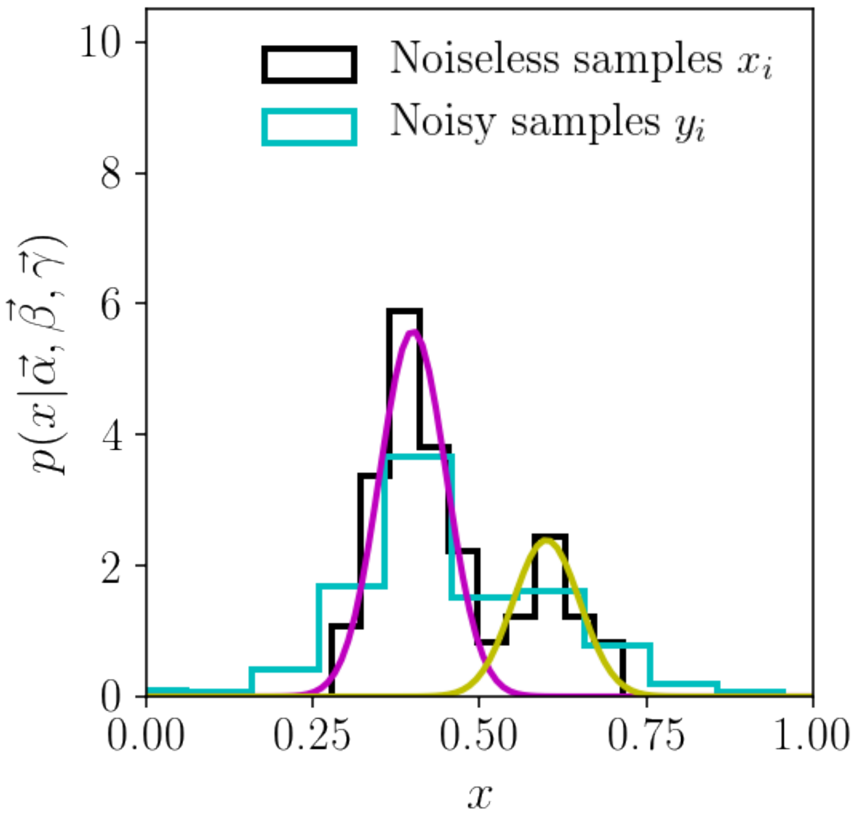

- Mixture model for density of true x's (which are latent parameters)

$$p(x_i|\vec{\alpha}, \vec{\beta}, \vec{\gamma}) = \sum_{b=1}^{B} \alpha_b \mathcal{N}(x_i|\beta_b, \gamma_b^2)$$

- Observed noisy y's: $p(y_i|x_i, \sigma_i) = \mathcal{N}(y_i|x_i, \sigma_i^2)$

- Posterior distribution (=deconvolution!)

$$p(\vec{f}, \vec{\beta}, \vec{\gamma}|\{y_i, \sigma_i\}) \propto p(\vec{f}, \vec{\beta}, \vec{\gamma}) \prod_{i=1}^{N} \sum_{b=1}^{B} f_b \int \mathrm{d}x_i \mathcal{N}(x_i|\beta_b, \gamma_b^2) \mathcal{N}(y_i|x_i, \sigma_i^2)$$




See http://ixkael.com for tutorials and code for Bayesian hierarchical models, uncertainty shrinkage, selection effects, etc

# Gaia DR1 color-magnitude diagram

*Leistedt & Hogg, ApJ 2017 (arXiv:1703.08112)*

- Data: Gaia TGAS cross-matched with APASS.
- Method: full hierarchical inference via Gibbs sampling.

*Anderson, Hogg, Leistedt, Price-Whelan, Bovy, ApJ 2017 (arXiv:1706.05055)*

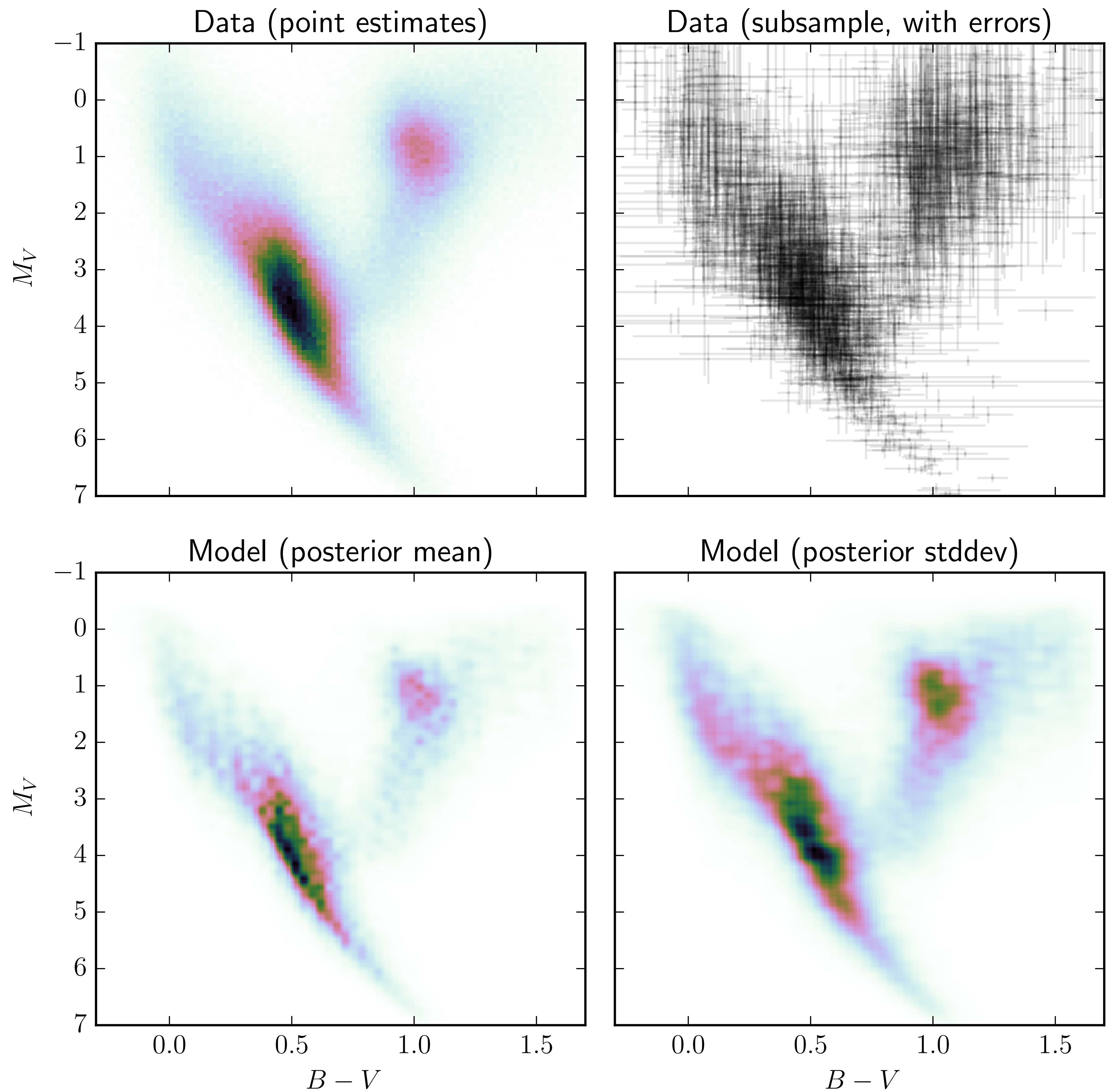
- Data: Gaia TGAS cross-matched with 2MASS.
- Method: extreme deconvolution and empirical Bayes.

# Full CMD hierarchical model

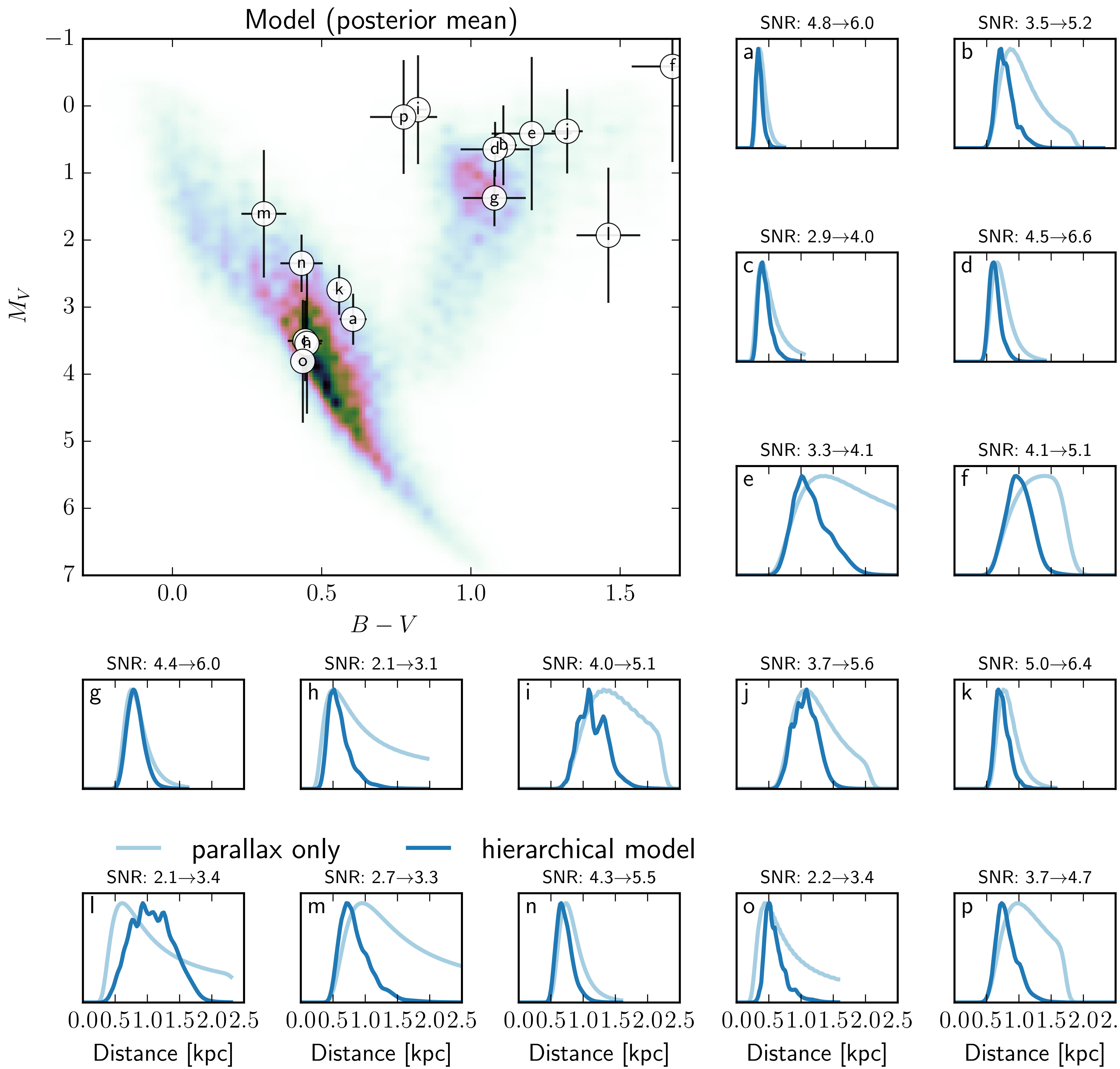
*Leistedt & Hogg, ApJ 2017 (arXiv:1703.08112)*

- Instead of using rigid stellar models, we will use all of the data (at all SNR) to construct a model of the color-magnitude diagram including all magnitude and color information and marginalizing over uncertainties.
- Mixture model: $p(M, C) = \sum_b \alpha_b \mathcal{N}(\vec{\mu}_b, \Sigma_b)$
- MCMC with Gibbs sampling. 3D dust fixed.
  Bins + distances marginalized over via sampling.
  True color + magnitude analytically marginalized over.

# Results: error-deconvolved HRD

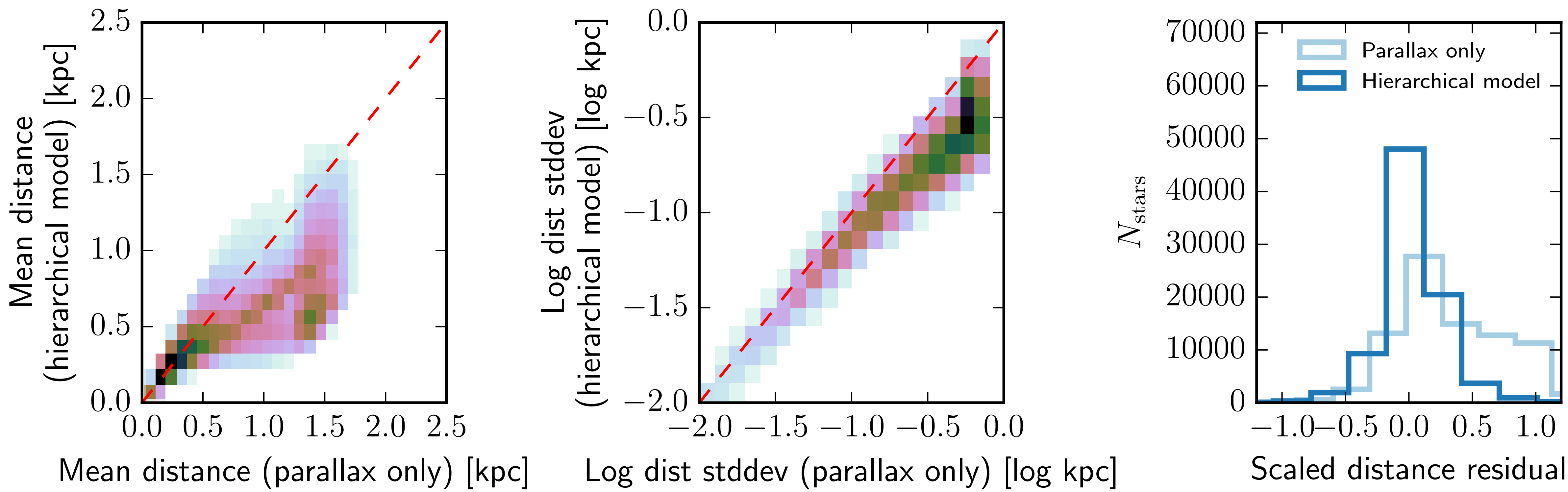


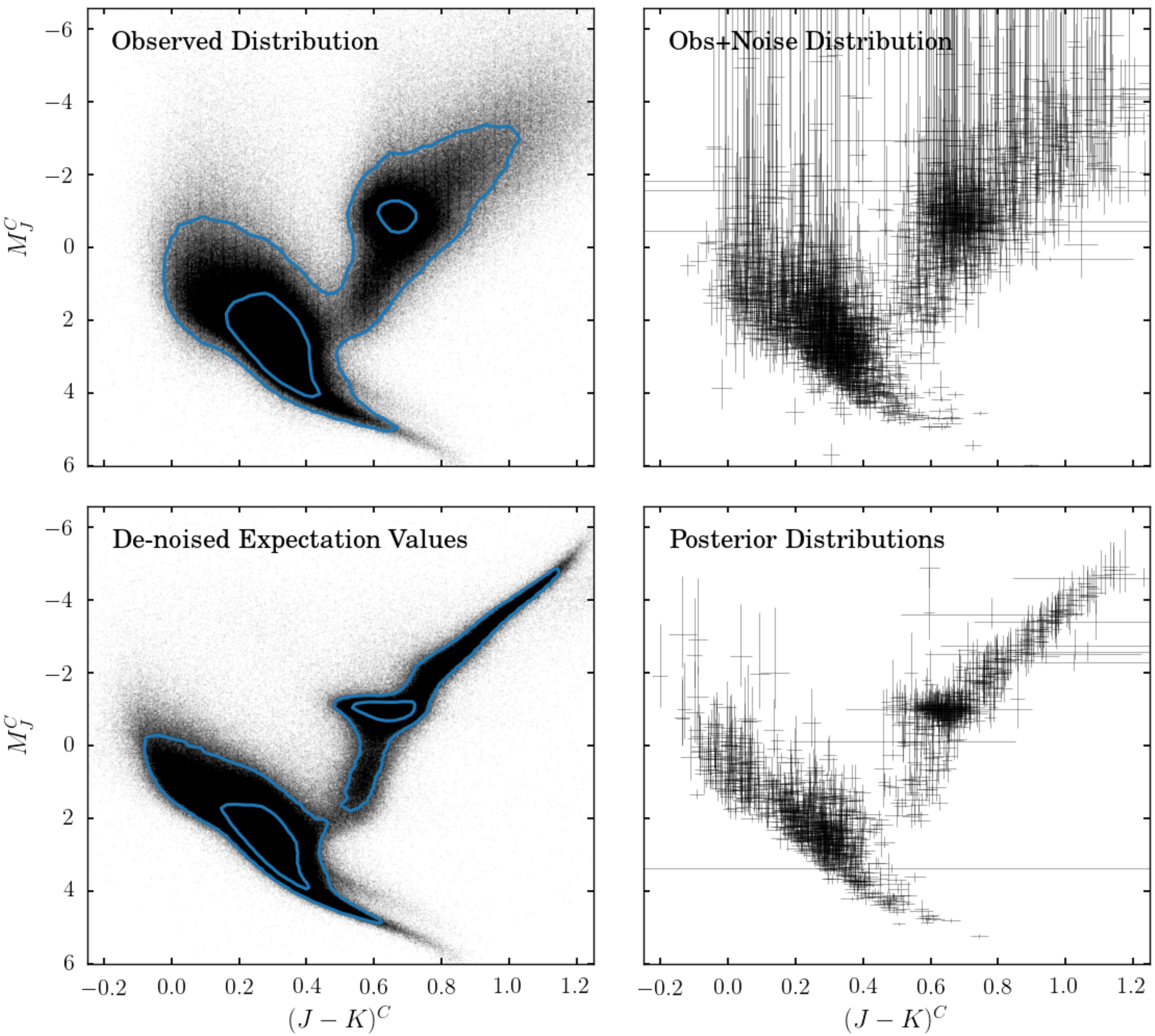
# Hierarchical uncertainty shrinkage



# Hierarchical uncertainty shrinkage

- Natural consequence of hierarchical models: the inferred population distributions act as priors on the internal variables.
- We constructed a color-magnitude diagram directly from the data, constraining the true color + absolute magnitude of each object, resulting in tighter constraints on the distances.

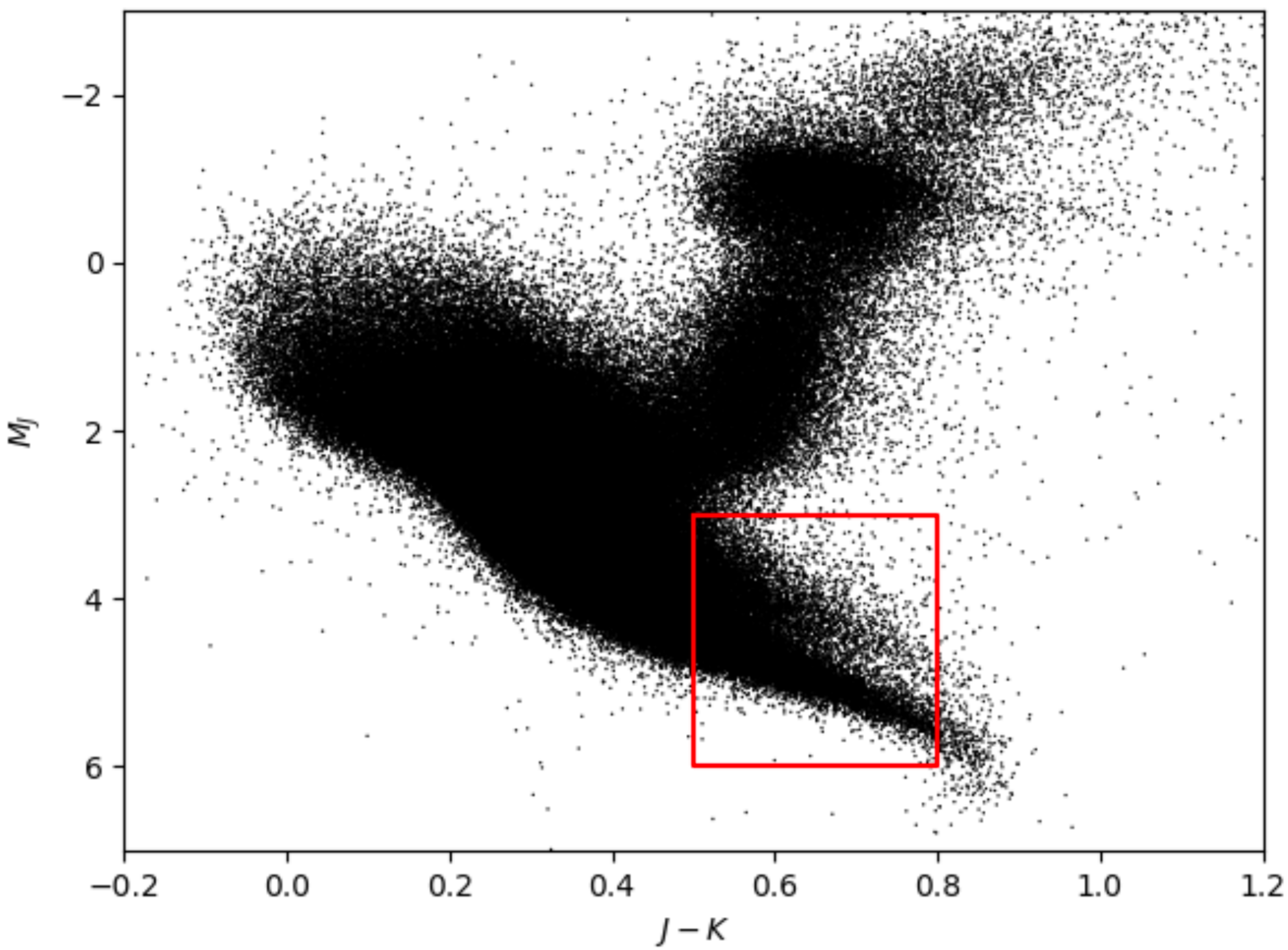





Anderson et al

# Evidence for double and triple sequences

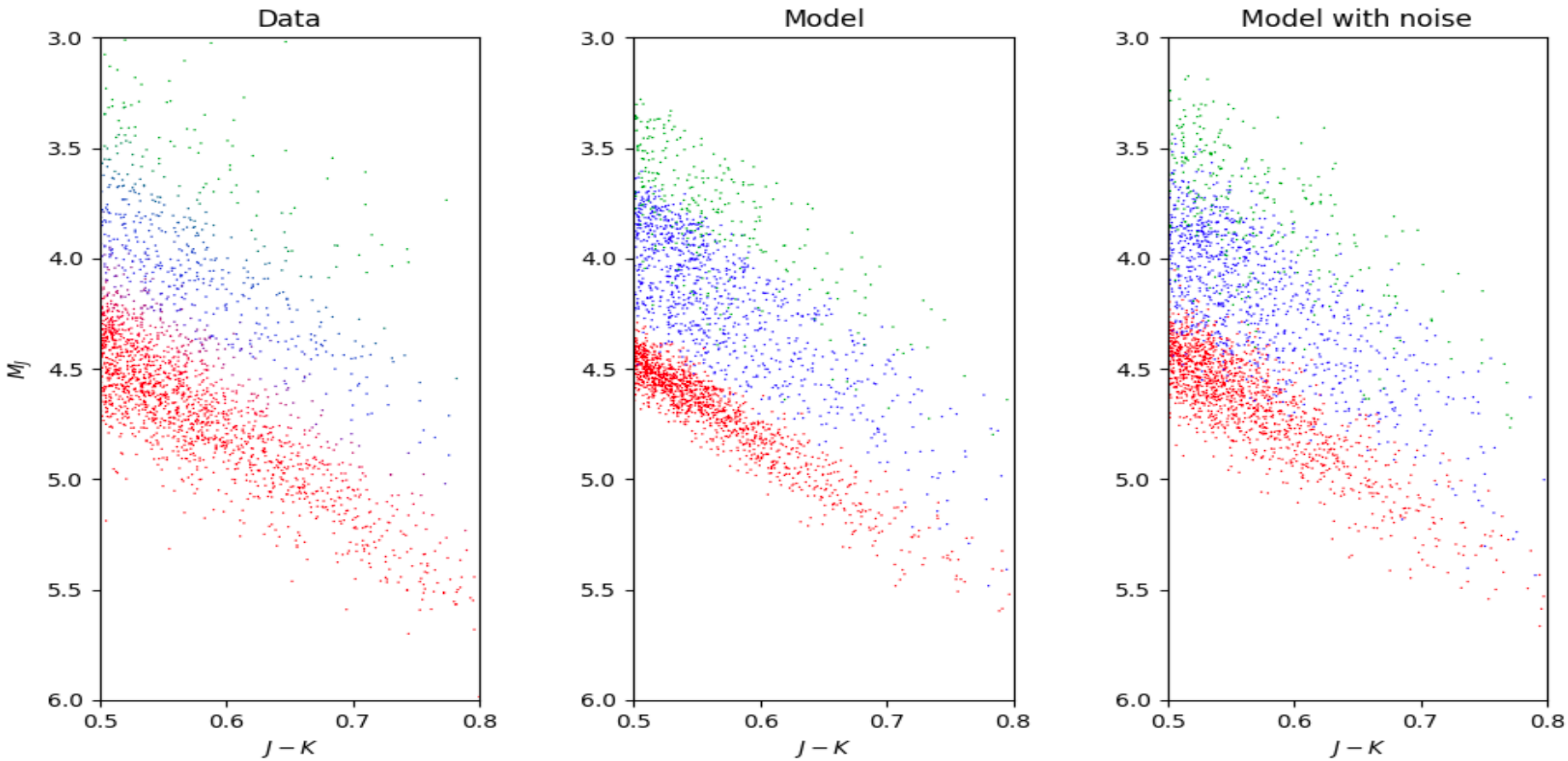
Preliminary work by Axel Widmark (Stockholm) with D. Hogg




Gaia TGAS+2MASS. Joint fit to CMD with singles & doubles.

# Evidence for double and triple sequences

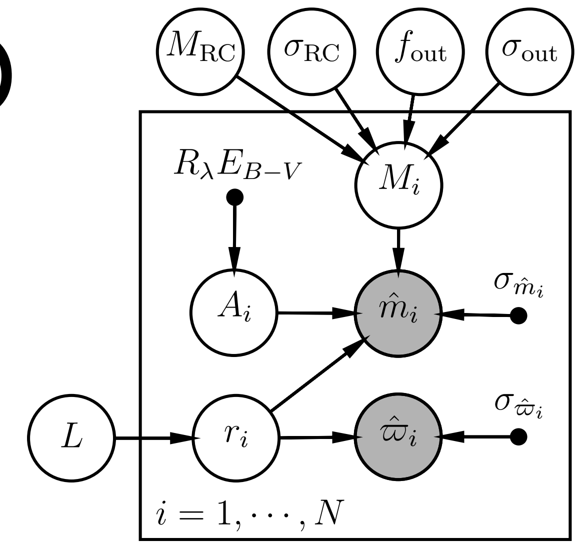
Preliminary work by Axel Widmark (Stockholm) with D. Hogg



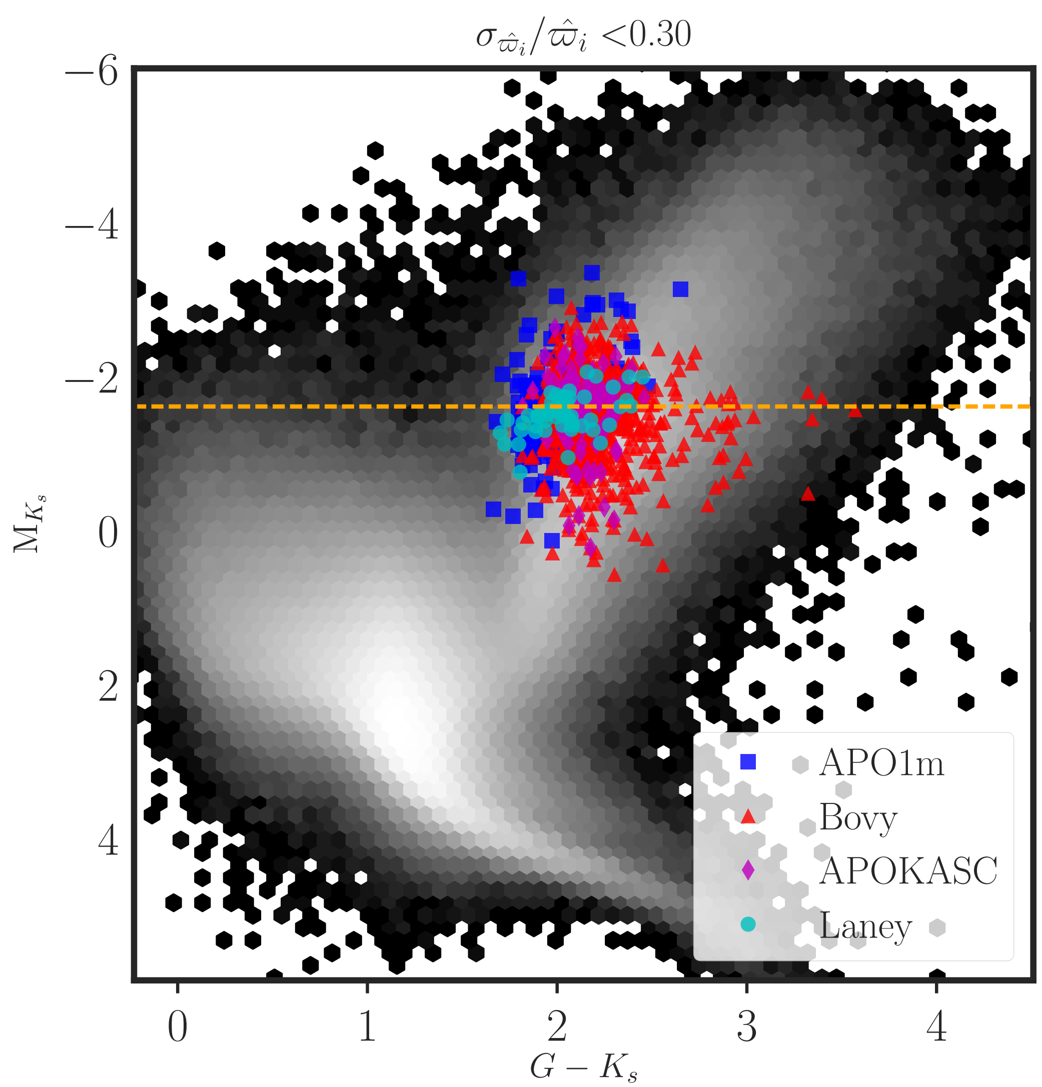

Data scatter well explained by unresolved binaries & triples.
Next steps: classification and connection to physical models.

# Calibration of the red-clump

Hawkins, Leistedt, Body & Hogg, MNRAS 2017 (arXiv:1705.08988)




Hierarchical modeling: Gaussian + outliers, marginalizing of dust, parallaxes, observed magnitudes.



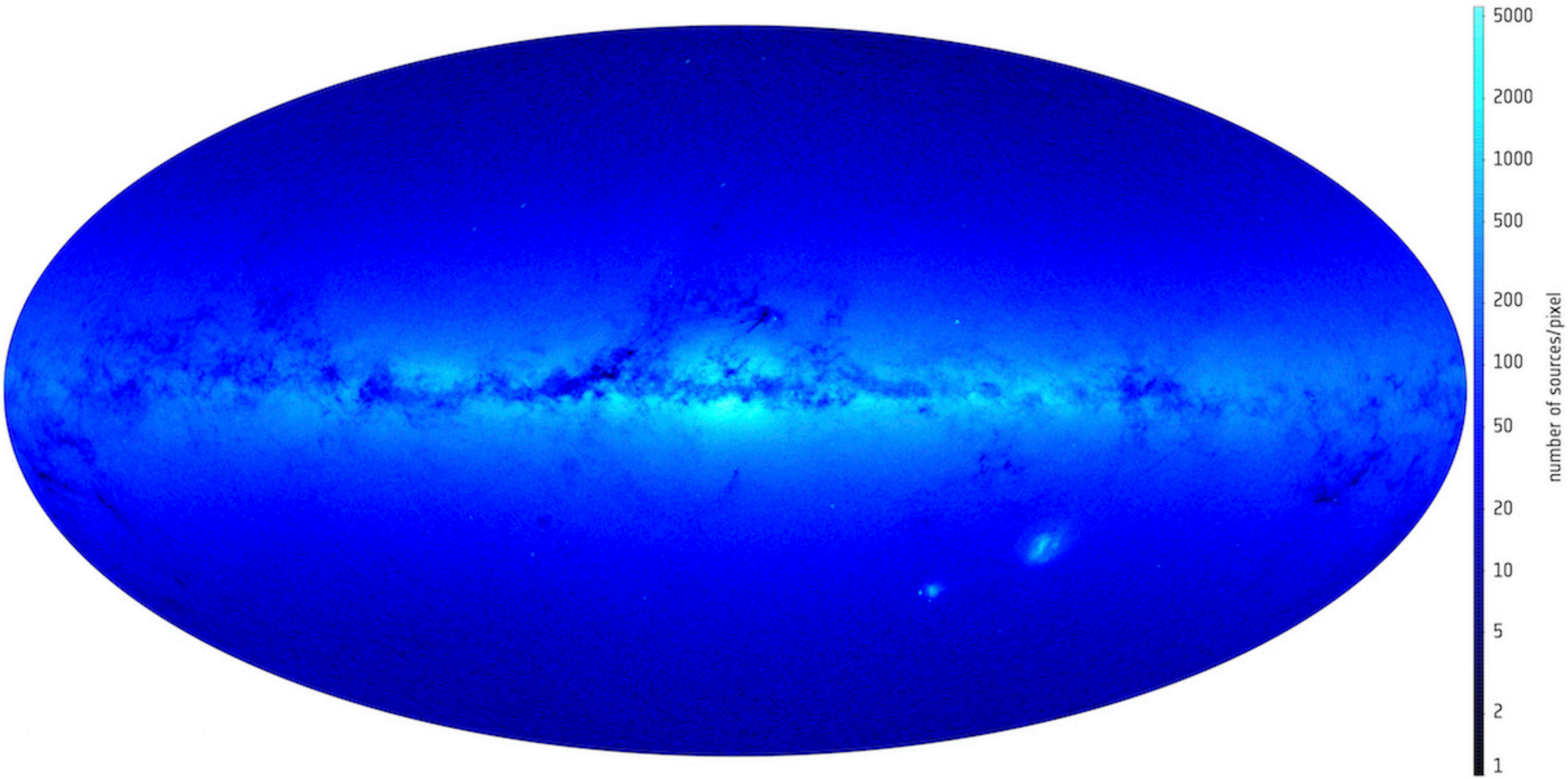

**RC absolute magnitude:**

K band: −1.61± 0.01 mag
G band: 0.44±0.01 mag
J band: −0.93±0.01 mag
H band: −1.46±0.01 mag
W1 band: −1.68±0.02 mag
W2 band: −1.69±0.02 mag
W3 band: −1.67±0.02 mag
W4 band: -1.76±0.01 mag

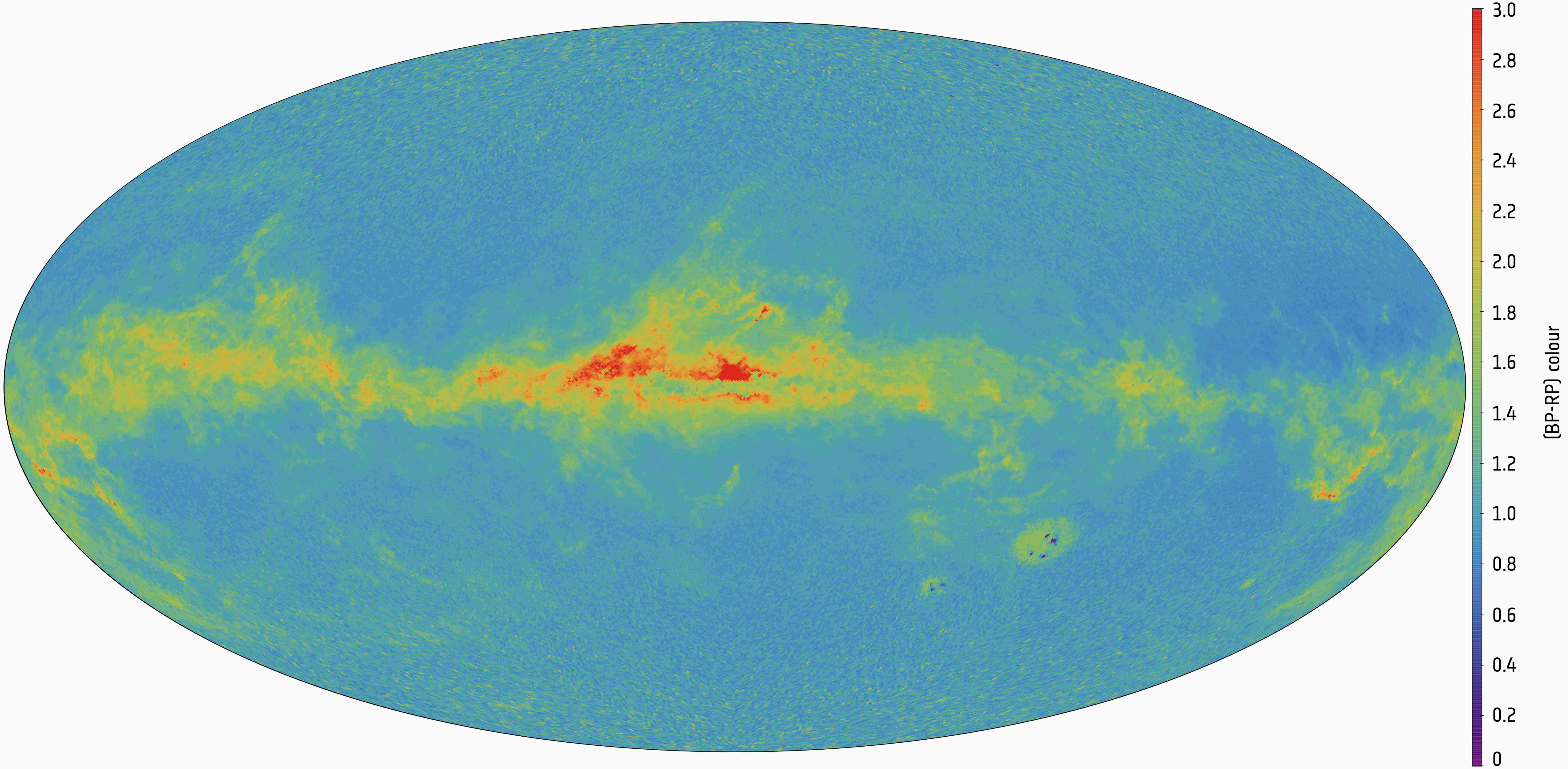
Intrinsic dispersion ~0.17±0.03 mag

Distance precision ~8%

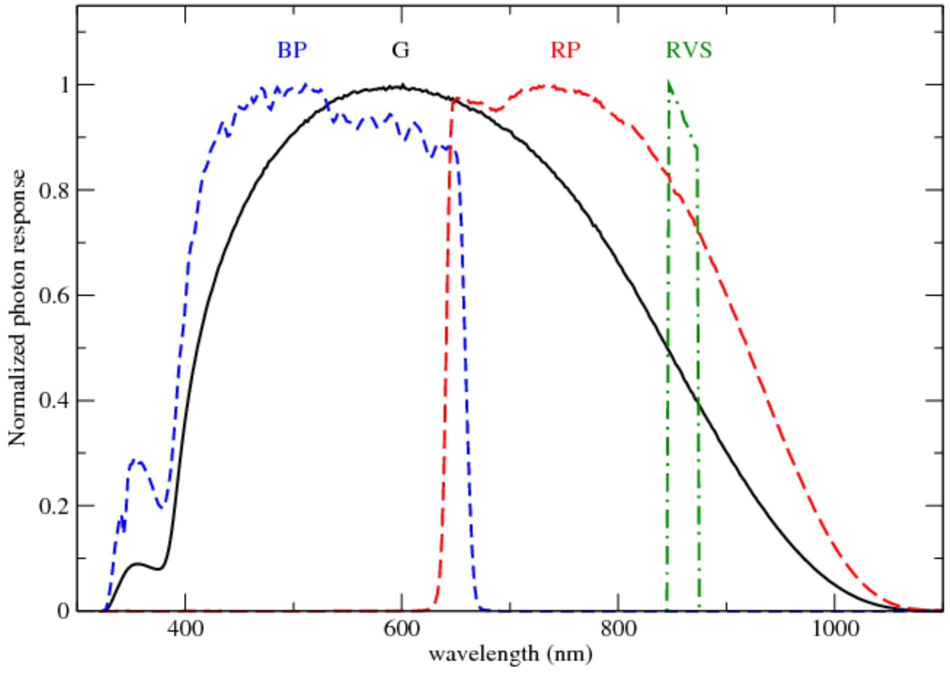
# Gaia DR2 (04/2018)



# Gaia DR2 (04/2018)



# Gaia DR2 (04/2018)

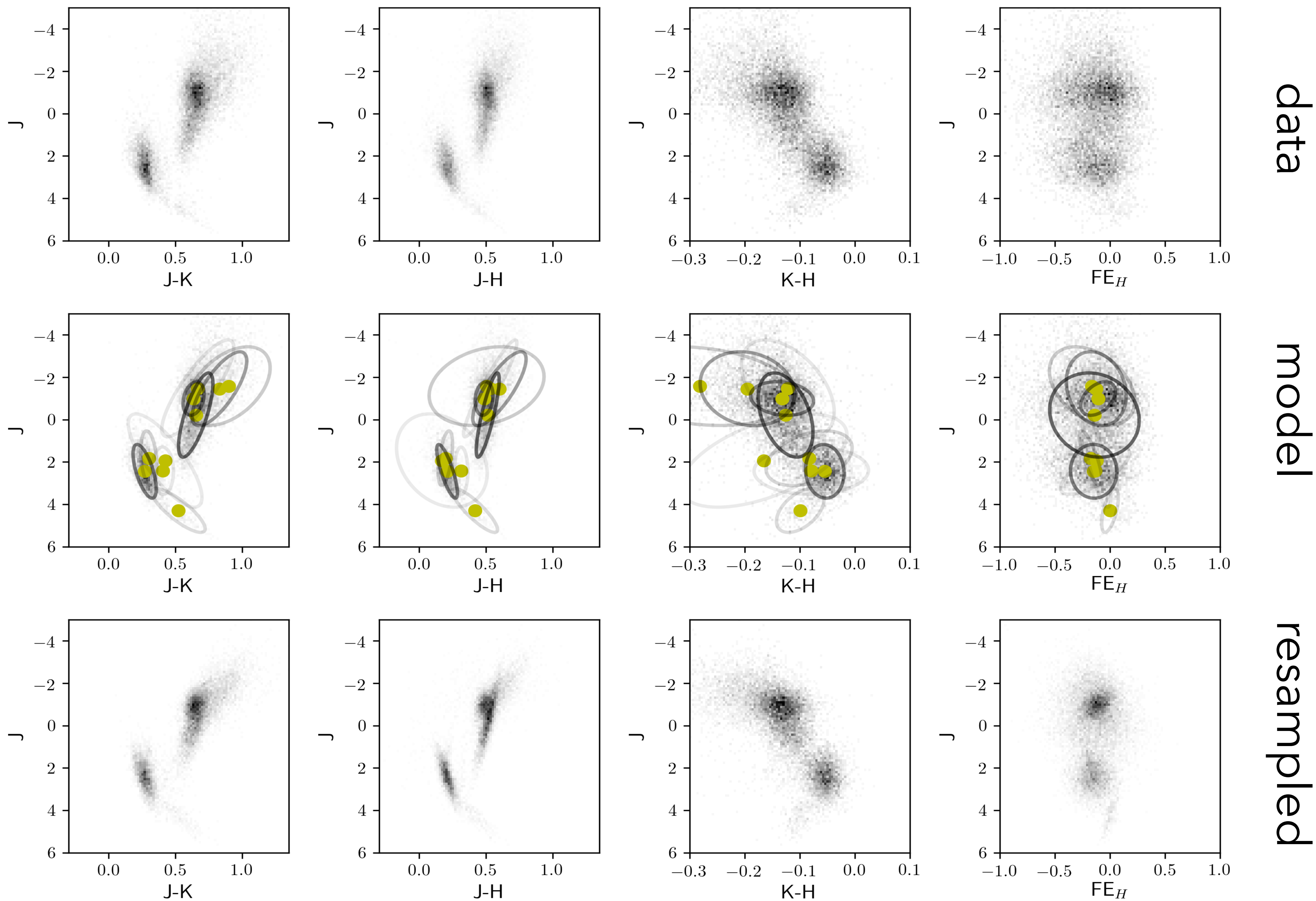



G + BR + RP magnitudes
+ 2MASS, WISE, etc
+ parallaxes/proper motions
= deep dynamic multi-color view of the Galaxy.

<u>Projects we will be ready to do:</u>

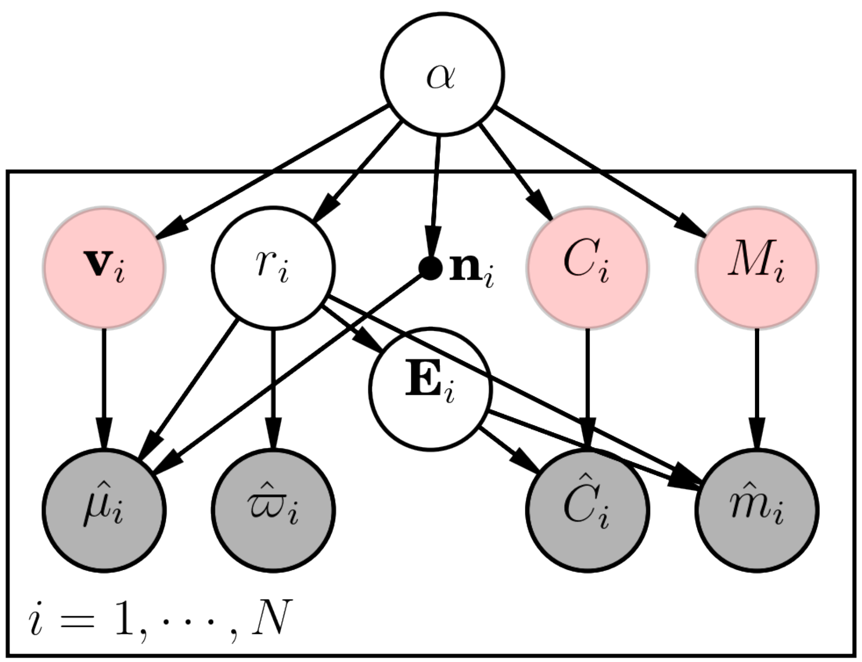
- Multicolor color-magnitude diagram
- Improved distance estimates using all the information
- Detailed 3D dust map directly only from Gaia data
- Metallicity map via transfer from RAVE/APOGEE

# Multicolor CMD (Gaia TGAS+2MASS, preliminary)




Efficient inference: numerical parallax marginalization, tensorflow SGD

# Full Gaia HPM




| | |
|---|---|
| $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_B)$ | All parameters of the mixture model |
| $\boldsymbol{\alpha}_b = (f_b, \boldsymbol{\xi}_b, \Sigma_b)$ | Parameters of the $b$ Gaussian of the mixture |
| $i$ | Index of the $i$th star |
| $\boldsymbol{n}_i = (\alpha_i, \delta_i)$ | True/observed angular position |
| $r_i$ | True distance |
| $\boldsymbol{v}_i = (v_{x,i}, v_{y,i}, v_{z,i})$ | True 3D cartesian velocity |
| $\hat{\boldsymbol{\mu}}_i = (\mu_{\alpha,i}, \mu_{\delta,i})$ | Observed proper motion |
| $\hat{\varpi}_i$ | Observed parallax |
| $\boldsymbol{E}_i \to E_{m_i}, E_{C_i}$ | True magnitude/color extinction at distance $r_i$ |
| $C_i, \hat{C}_i$ | True and observed color |
| $M_i$ | True absolute magnitude |
| $\hat{m}_i$ | Observed apparent magnitude |

Infer the distributions from the data (here in 8D)

Analytic or numerical marginalization of latent parameters.

GMM model: $[\boldsymbol{v}\ \boldsymbol{n}\ r\ C\ M]^T |\ \boldsymbol{\alpha} \sim \sum_{b=1}^{B} f_b\, \mathcal{N}^{8\mathrm{D}}(\boldsymbol{\xi}_b; \Sigma_b).$

Technology: stochastic gradients, Tensorflow, etc

# Summary

Gaia: exciting data set, but computationally challenging.

We developed inference techniques and data-driven models for fully & correctly exploiting all of the data.

Gaia DR1: high-precision color-magnitude diagrams, binary/triple sequences, improved stellar distances

Gaia DR2 (April 2018): 3D reconstruction of stellar density, dust, and velocities.

Codes/experiments public on github.com/ixkael