UMass Bayesian Inference Engine

Martin D. Weinberg
UMass Astronomy
weinberg@astro.umass.edu

Current team

- MDW (Astro)
- Neal Katz (Astro)
- Michael Lavine (Math)
- Houjun Mo (Astro)
- Eliot Moss (CS)
- Byn Choi (CS)
- Joerg Colberg (Astro)
- Mark Fardal (Astro)
- Ilsang Yoon (Astro)
- Lu Yu (Astro)

Motivation for the Bayesian Inference Engine

- Multi-terabyte catalogs (2MASS, SDSS, GOODS, . . .)
- Bayesian approach
  - Incorporate data from multiple catalogs
  - Merge data with different attributes
- Computational solution to the inference problem
  - MCMC algorithms
  - Particle filter
- Current packages—Bayespack, BUGS, S-Plus, R
  - Not production oriented
  - Although good for proof of concept

Motivation for the Bayesian Inference Engine

- Multi-terabyte catalogs (2MASS, SDSS, GOODS, . . .)
- Bayesian approach
  - Incorporate data from multiple catalogs
  - Merge data with different attributes
- Computational solution to the inference problem
  - MCMC algorithms
  - Particle filter
- Current packages—Bayespack, BUGS, S-Plus, R
  - Not production oriented
  - Although good for proof of concept

Motivation for the Bayesian Inference Engine

- Multi-terabyte catalogs (2MASS, SDSS, GOODS, . . .)
- Bayesian approach
  - Incorporate data from multiple catalogs
  - Merge data with different attributes
- Computational solution to the inference problem
  - MCMC algorithms
  - Particle filter
- Current packages—Bayespack, BUGS, S-Plus, R
  - Not production oriented
  - Although good for proof of concept

Motivation for the Bayesian Inference Engine

- Multi-terabyte catalogs (2MASS, SDSS, GOODS, . . .)
- Bayesian approach
  - Incorporate data from multiple catalogs
  - Merge data with different attributes
- Computational solution to the inference problem
  - MCMC algorithms
  - Particle filter
- Current packages—Bayespack, BUGS, S-Plus, R
  - Not production oriented
  - Although good for proof of concept

BIE: Current Features

1. Do it fast!
2. Do it better!
3. Do it right!

BIE: Current Features

1. Do it fast!
   Apply advanced computational techniques
   - Fully parallel implementation (MPI)
   - Object-oriented, consumer-producer pipes
   - C++ library with front-end parser
   - Full serialization and persistence
     - Checkpointing
     - Reuse of posterior distributions
   - Visualization tools
2. Do it better!
3. Do it right!

BIE: Current Features

1. Do it fast!
2. Do it better!
3. Do it right!

BIE: Current Features

1. Do it fast!
2. Do it better!
   Platform for future development and statistical research
   - Convergence and posterior characterization
     - Covariance analysis, Laplace approximation
     - Metric Ball Trees (KDE)
   - New computational techniques (Bayes factors)
   - Reversible jump for mixture models
   - New MCMC algorithms
     - Empirical hierarchical priors
     - Tempered differential evolution
3. Do it right!

BIE: Current Features

1. Do it fast!
2. Do it better!
3. Do it right!

BIE: Current Features

1. Do it fast!
2. Do it better!
3. Do it right!

   Perform meaningful inference & hypothesis testing on survey datasets

Provide inference beyond χ^2!

- Encourage rigorous model selection
  - Bayes factor, evidence computation
- Mixture models (multiple components)
  - Dimension switching: reversible jump
- General hypothesis testing
  - Without nested models
  - Test complex hypotheses
- Correlations from pooled posterior distributions
  $\Rightarrow$ NOT scatter diagrams

Killer applications

1. Star count/isochrone analysis
2. Semi-analytic models (SAMS)
   - BIE-SAM
3. Galaxy image analysis
   - Galaxy photometric attributes
   - GALaxy PHotometric ATtributes
   - GALPHAT

Semi-analytic models

- BIE-SAM: incorporates features from major groups
- Current practice
  - Fix some parameters
  - Adjust others by hand to fit observations
  - Chi-by-eye
  - Summary data: luminosity & mass function, Tully-Fisher relation
- Problem: no confidence regions⇒CANNOT achieve goal of rejecting phenomenological models
- Posterior distribution is complex

Semi-analytic models

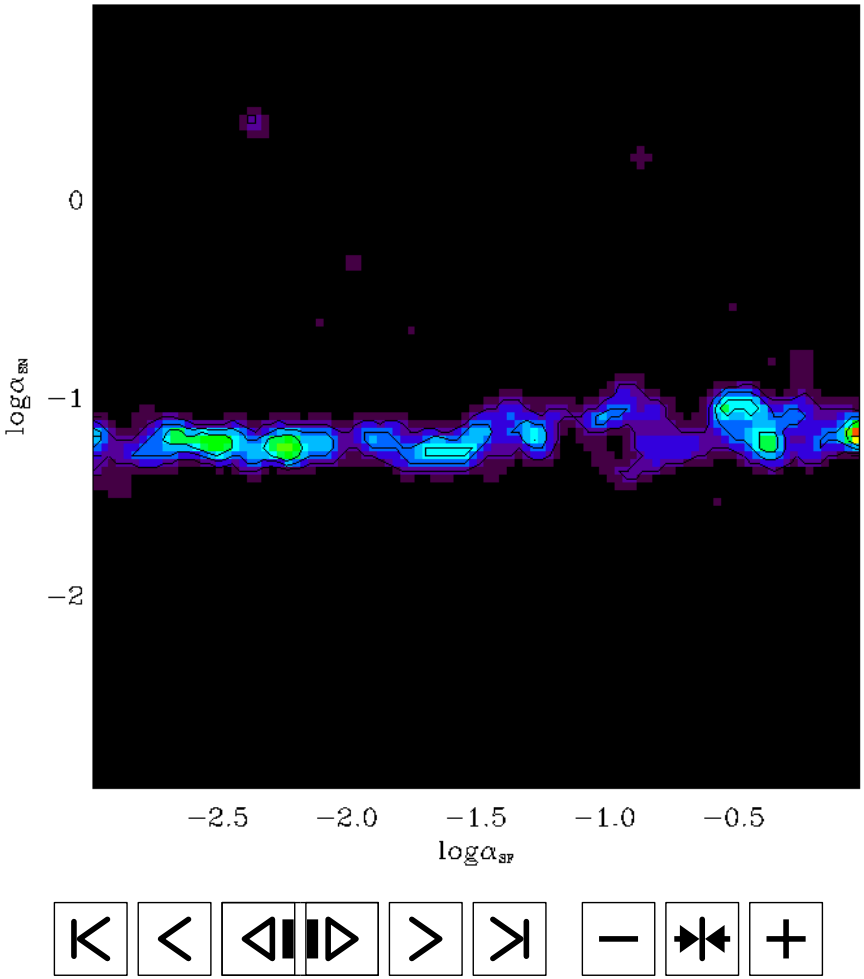
- BIE-SAM: incorporates features from major groups
- Current practice
  - Fix some parameters
  - Adjust others by hand to fit observations
  - Chi-by-eye
  - Summary data: luminosity & mass function, Tully-Fisher relation
- Problem: no confidence regions⇒CANNOT achieve goal of rejecting phenomenological models
- Posterior distribution is complex

Semi-analytic models

- BIE-SAM: incorporates features from major groups
- Current practice
  - Fix some parameters
  - Adjust others by hand to fit observations
  - Chi-by-eye
  - Summary data: luminosity & mass function, Tully-Fisher relation
- Problem: no confidence regions⇒CANNOT achieve goal of rejecting phenomenological models
- Posterior distribution is complex

Posterior madness: why we need fancy MCMC

- Example: 13 parameter model given galaxy mass function



- BIE-SAM: all but 3 out of 13 components marginalized
- Each panel: star formation (SF) vs supernova feedback efficiency
- Steps in gas surface density threshold for SF: large to small
- Bell et al. (2003) K-band galaxy mass function

GALPHAT image analysis: main features

- Multicomponent modeling (typically 12 parameters)
- Background estimation or specification
- Adaptive integration, optimized with two-dimensional interpolation on cumulative distributions
- Rotation by FFT shear algorithm

GALPHAT: main features

- Scientific goals
  1. Evaluation of galaxy evolution theories (model selection)
  2. Look for correlations between inferred parameters (knowledge discovery)
- Current projects
  - Bulge/disk ratios from 2MASS & SDSS
  - Higher redshift (GOODS, GEMS): evolution of bulge/disk ratio, correlation with environment
  - Hypothesis testing with full posterior probabilities

GALPHAT: main features

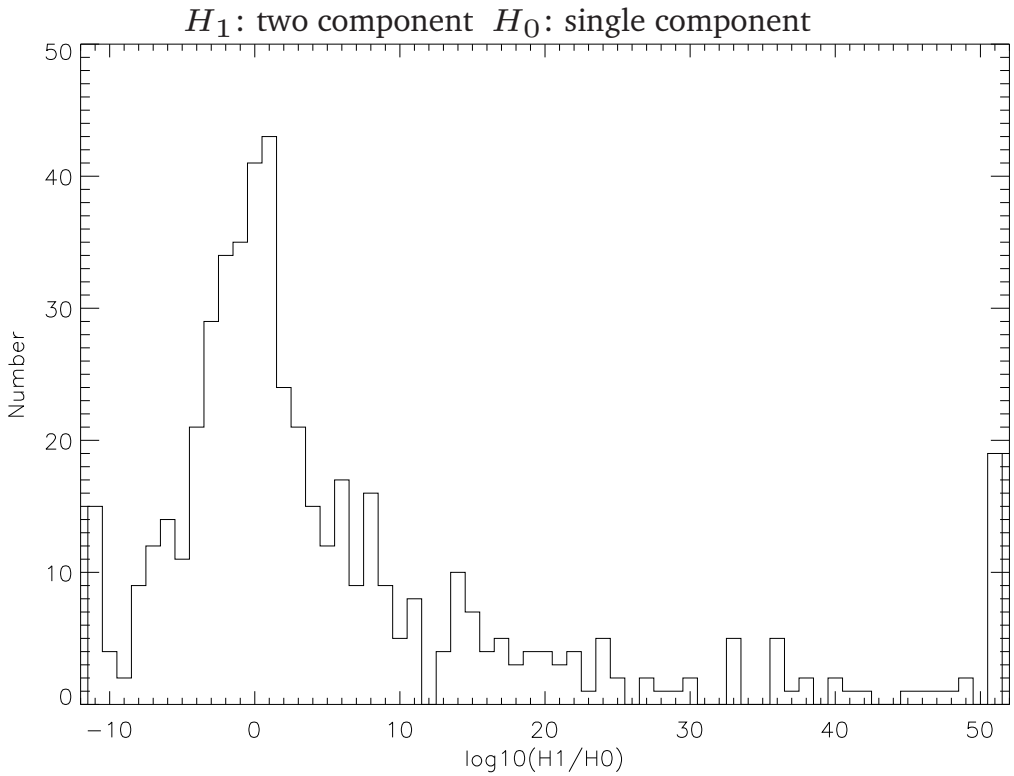
- Scientific goals
  1. Evaluation of galaxy evolution theories (model selection)
  2. Look for correlations between inferred parameters (knowledge discovery)
- Current projects
  - Bulge/disk ratios from 2MASS & SDSS
  - Higher redshift (GOODS, GEMS): evolution of bulge/disk ratio, correlation with environment
  - Hypothesis testing with full posterior probabilities

GALPHAT: results

- 510 2MASS galaxies in SDSS ($8.1 < K < 10.23$)
- Single and double component Sérsic models
- 25000 converged MCMC samples for every galaxy

GALPHAT: results

- 510 2MASS galaxies in SDSS ($8.1 < K < 10.23$)
- Single and double component Sérsic models
- 25000 converged MCMC samples for every galaxy



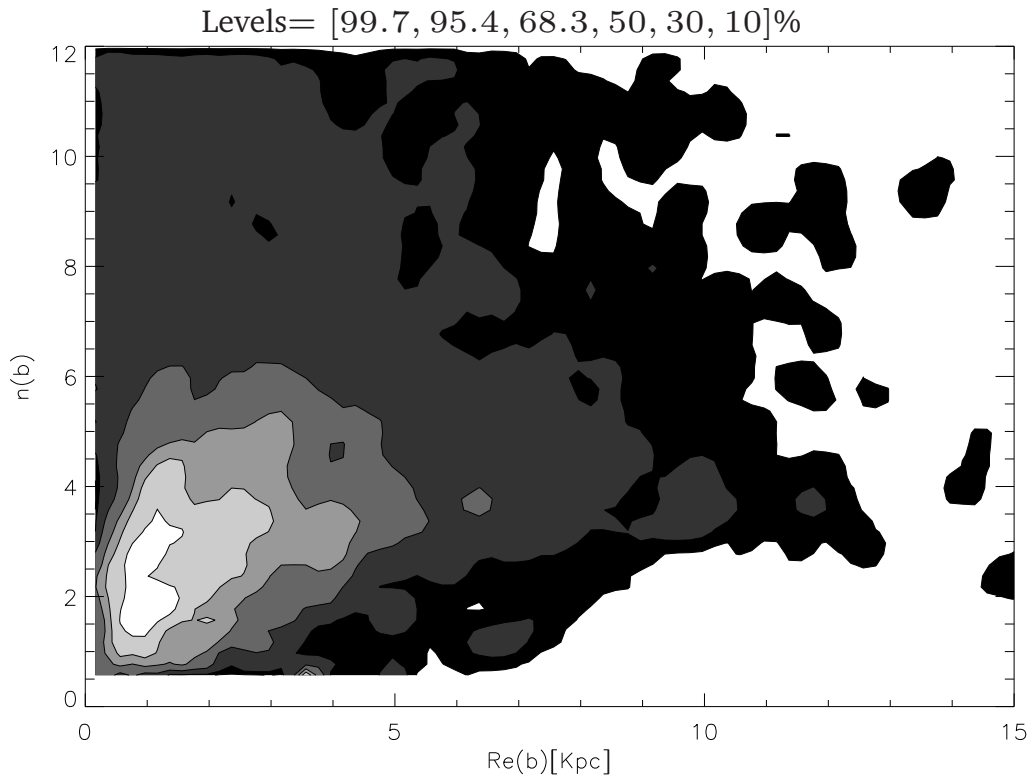
Bayesian evidence ratio

GALPHAT: results

- Classification:
  - $H_1/H_0 > 1 \Rightarrow$ bulge + disk
  - $H_1/H_0 < 1 \Rightarrow$ single Sérsic (e.g. elliptical)

GALPHAT: results

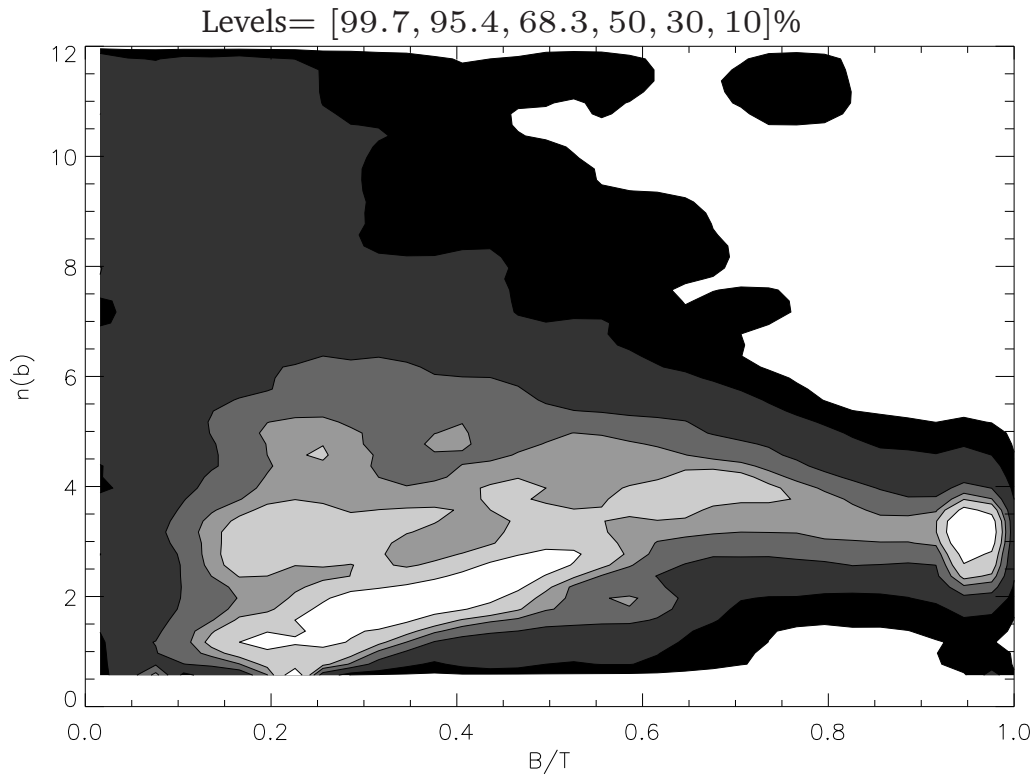
- Classification:
  - $H_1/H_0 > 1 \Rightarrow$ bulge + disk
  - $H_1/H_0 < 1 \Rightarrow$ single Sérsic (e.g. elliptical)



Bulge radius vs Sérsic index

GALPHAT: results

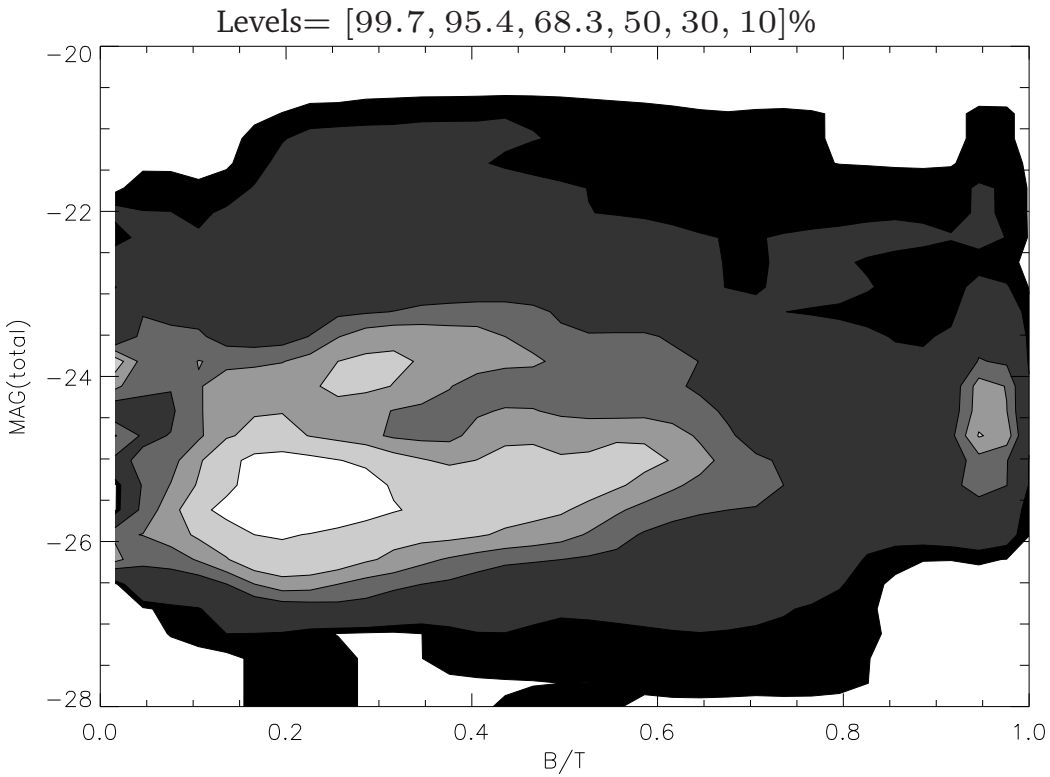
- Classification:
  - $H_1/H_0 > 1 \Rightarrow$ bulge + disk
  - $H_1/H_0 < 1 \Rightarrow$ single Sérsic (e.g. elliptical)



Bulge/Total vs Sérsic index

GALPHAT: results

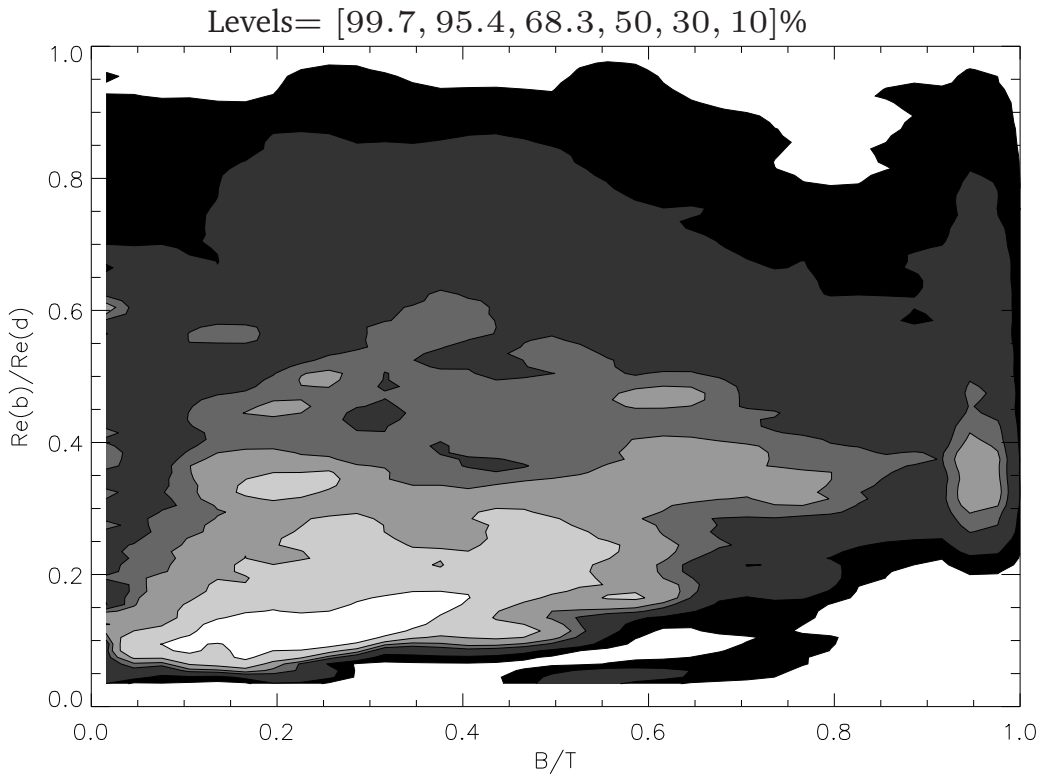
- Classification:
  - $H_1/H_0 > 1 \Rightarrow$ bulge + disk
  - $H_1/H_0 < 1 \Rightarrow$ single Sérsic (e.g. elliptical)



Bulge/Total vs Mag

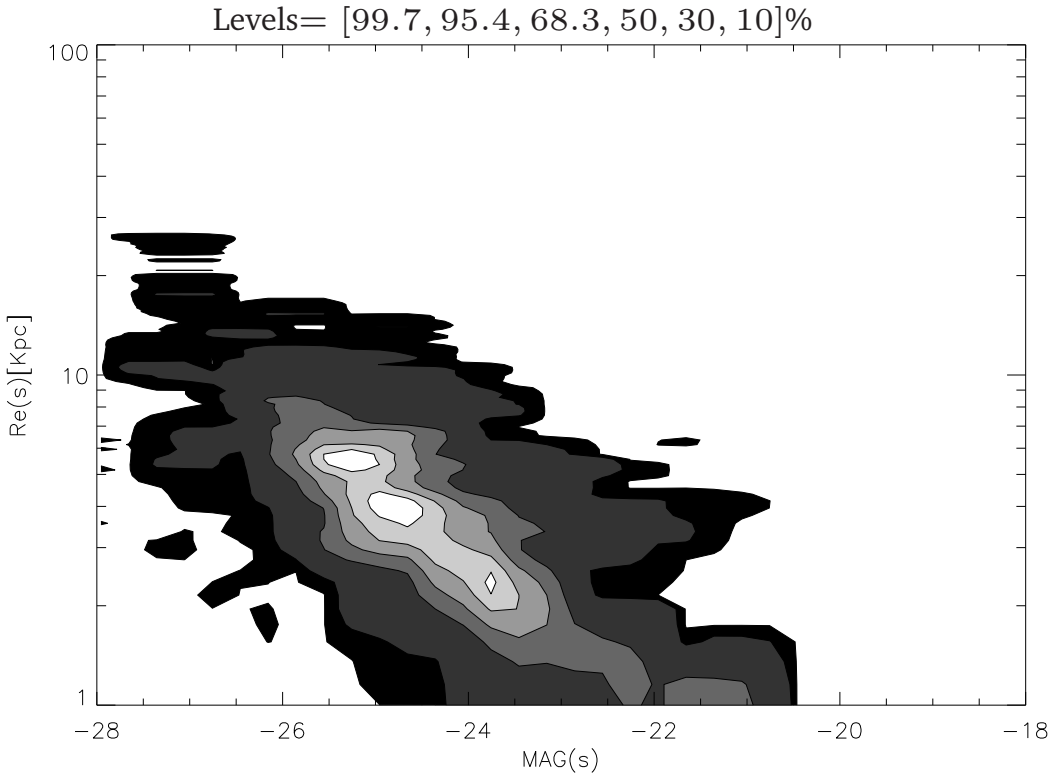
GALPHAT: results

- Classification:
  - $H_1/H_0 > 1 \Rightarrow$ bulge + disk
  - $H_1/H_0 < 1 \Rightarrow$ single Sérsic (e.g. elliptical)



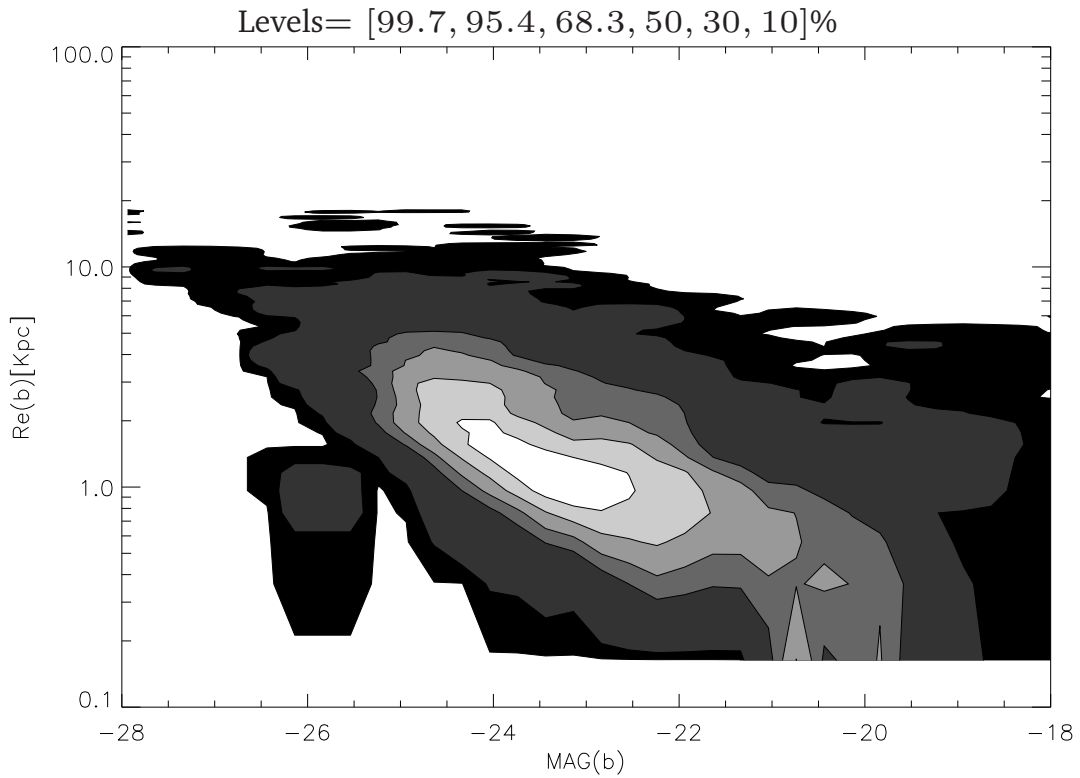
Bulge/Total vs size ratio

Size correlations



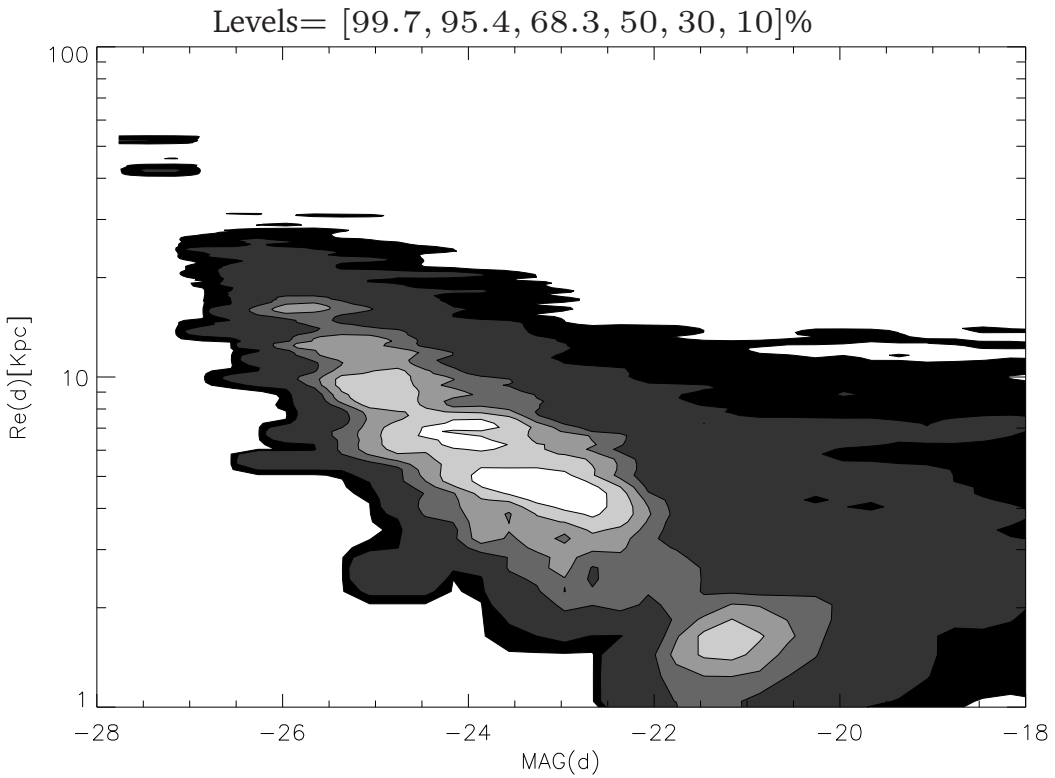
Mag vs R_e for single component

Size correlations



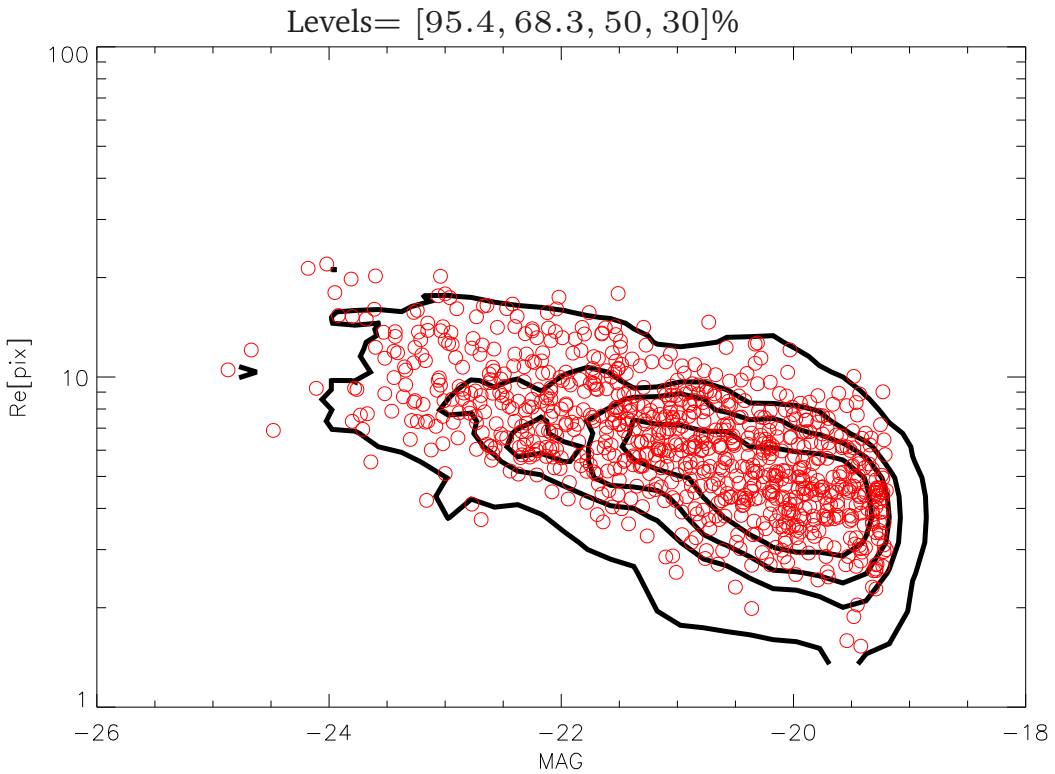
Mag vs R_e for bulges (two component)

Size correlations



Mag vs R_e for disks (two component)

Synthetic image tests



Mag vs R_e
Input (points) and inferred posterior (contours)

GALPHAT: summary

- Quantitative inference of galaxy parameters
  - Classification with confidence!
  - Quantification of morphological trends
- Easily extended to more general families
- Ultimately: non-parametric families
  $\Rightarrow$ knowledge discovery

GALPHAT: summary

- Quantitative inference of galaxy parameters
  - Classification with confidence!
  - Quantification of morphological trends
- Easily extended to more general families
- Ultimately: non-parametric families
  $\Rightarrow$ knowledge discovery

GALPHAT: summary

- Quantitative inference of galaxy parameters
  - Classification with confidence!
  - Quantification of morphological trends
- Easily extended to more general families
- Ultimately: non-parametric families
  $\Rightarrow$ knowledge discovery

Current

- Listed on Astro Stat web site
- Project web site: `www.astro.umass.edu/~weinberg/bie`
- 2010 release including persistence and standalones
- Interim releases now, contact me!

Current status and Future projects

Current

- Listed on Astro Stat web site
- Project web site: `www.astro.umass.edu/~weinberg/bie`
- 2010 release including persistence and standalones
- Interim releases now, contact me!

Future

- GALPHAT: CUDA/OpenCL design
- Kinematic/photometric mass estimation ("rotation curves")

Current status and Future projects

Current

- Listed on Astro Stat web site
- Project web site: `www.astro.umass.edu/~weinberg/bie`
- 2010 release including persistence and standalones
- Interim releases now, contact me!

Future

- GALPHAT: CUDA/OpenCL design
- Kinematic/photometric mass estimation ("rotation curves")

Done!