Bayesian constraints on the physics of galaxy formation

Yu Lu
KIPAC/Stanford

Houjun Mo, Neal Katz, Martin Weinberg (UMass)

MADAI workshop, SAMSI, 07/29/2013
Dark Energy
72%  Stays uniform

Dark Matter
23%  Clumps with time

Visible
Baryons
4.6%
hydrodynamical simulation

- numerically solve equations govern the nonlinear evolution of dark matter and baryons.
- limited dynamical range
- sub-resolution recipes
- expensive computation
- hard to explore model parameter space

Springel & Hernquist 2003
Semi-Analytic Model (SAM)

- Phenomenological model, parameterize important processes
- Processes in SAM:
  - Dark matter halos: distribution, growth
  - Hot gas distribution
  - Radiative cooling
  - Star formation and supernova feedback
  - Galaxy merger
  - AGN, reionization, environmental effects…
- Monte Carlo realization
Semi-Analytic Model (SAM)

- Phenomenological model, parameterize important processes

Processes in SAM:
- Dark matter halos: distribution, growth
- Hot gas distribution
- Radiative cooling
- Star formation and supernova feedback
- Galaxy merger
- AGN, reionization, environmental effects...

- Monte Carlo realization
Semi-Analytic Model (SAM)

- Phenomenological model, parameterize important processes
- Processes in SAM:
  - Dark matter halos: distribution, growth
  - Hot gas distribution
  - Radiative cooling
  - Star formation and supernova feedback
  - Galaxy merger
  - AGN, reionization, environmental effects…
- Monte Carlo realization
Semi-Analytic Model (SAM)

- Phenomenological model, parameterize important processes
- Processes in SAM:
  - Dark matter halos: distribution, growth
  - Hot gas distribution
  - Radiative cooling
  - Star formation and supernova feedback
  - Galaxy merger
  - AGN, reionization, environmental effects…
- Monte Carlo realization
• SAM is a useful method -
  - Predictive: Luminosity function (stellar mass function), Tully-Fisher relation, colors, clustering, morphologies
  - Low computation cost
  - Model inference, hypothesis test
  - Complementary to simulations: explore model space

• The implementation is problematic -
  - Tweak parameters by hand, fit by eye
  - No systematic way for exploring model space
  - No rigorous way for model inference
  - No rigorous way for model test
SAM - as a problem of model inference

• We are given:
  ▶ a model (hypothesis),
  ▶ a plausible range for the parameters (prior),
  ▶ data.

• We ask for:
  ▶ the probability distribution of the model parameters that can explain the data (confidence range of model parameters - posterior),
  ▶ the degree of belief that the model is supported by the data,
  ▶ any robust predictions can be made.
Bayesian model inference

• Bayes theorem:  
  \[ p(\Theta | D) \propto p(\Theta) L(D | \Theta) \]

• Marginalized posterior:  
  \[
p(\Theta_m | D) = \int p(\Theta | D) d\Theta_n, \quad \Theta = \{ \Theta_m, \Theta_n \}
  \]

• Bayesian evidence (Occam's razor):  
  \[
p(D | M) = \int p(\Theta | M) L(D | \Theta, M) d\Theta
  \]

\[
\frac{p(M_1 | D)}{p(M_2 | D)} = \frac{p(M_1) p(D | M_1)}{p(M_2) p(D | M_2)}
\]

• Prediction:  
  \[
p(D' | D) = \int p(D' | \Theta) p(\Theta | D) d\Theta
  \]
Bayesian model inference

• Bayes theorem:
  \[ p(\Theta|D) \propto p(\Theta)L(D|\Theta) \]

• Marginalized posterior:
  \[ p(\Theta_m|D) = \int p(\Theta|D)\,d\Theta_n, \quad \Theta = \{\Theta_m, \Theta_n\} \]

• Bayesian evidence (Occam's razor):
  \[ p(D|M) = \int p(\Theta|M)L(D|\Theta, M)d\Theta \]
  \[ \frac{p(M_1|D)}{p(M_2|D)} = \frac{p(M_1)p(D|M_1)}{p(M_2)p(D|M_2)} \]

• Prediction:
  \[ p(D'|D) = \int p(D'|\Theta)p(\Theta|D)d\Theta \]
Bayesian model inference

- Bayes theorem:
  \[ p(\Theta | D) \propto p(\Theta) L(D | \Theta) \]

- Marginalized posterior:
  \[ p(\Theta_m | D) = \int p(\Theta | D) d\Theta_n, \quad \Theta = \{\Theta_m, \Theta_n\} \]

- Bayesian evidence (Occam's razor):
  \[ p(D | M) = \int p(\Theta | M) L(D | \Theta, M) d\Theta \]

\[ \frac{p(M_1 | D)}{p(M_2 | D)} = \frac{p(M_1) p(D | M_1)}{p(M_2) p(D | M_2)} \]

- Prediction:
  \[ p(D' | D) = \int p(D' | \Theta) p(\Theta | D) d\Theta \]
Bayesian model inference

• Bayes theorem: \[ p(\Theta|D) \propto p(\Theta)L(D|\Theta) \]

• Marginalized posterior:
\[ p(\Theta_m|D) = \int p(\Theta|D)d\Theta_n, \Theta = \{\Theta_m, \Theta_n\} \]

• Bayesian evidence (Occam's razor):
\[ p(D|M) = \int p(\Theta|M)L(D|\Theta,M)d\Theta \]
\[ \frac{p(M_1|D)}{p(M_2|D)} = \frac{p(M_1)p(D|M_1)}{p(M_2)p(D|M_2)} \]

• Prediction:
\[ p(D'|D) = \int p(D'|\Theta)p(\Theta|D)d\Theta \]
Bayesian model inference

• Bayes theorem: 
  \[ p(\Theta|D) \propto p(\Theta)L(D|\Theta) \]

• Marginalized posterior: 
  \[ p(\Theta_m|D) = \int p(\Theta|D)d\Theta_n, \Theta = \{\Theta_m, \Theta_n\} \]

• Bayesian evidence (Occam's razor): 
  \[ p(D|M) = \int p(\Theta|M)L(D|\Theta,M)d\Theta \]
  \[ \frac{p(M_1|D)}{p(M_2|D)} = \frac{p(M_1)p(D|M_1)}{p(M_2)p(D|M_2)} \]

• Prediction: 
  \[ p(D'|D) = \int p(D'|\Theta)p(\Theta|D) d\Theta \]
Metropolis–Hastings algorithm

\[ p(x) \]

\[ Q(x'; x_t) \]

\[ \alpha = \frac{p(x') Q(x_t; x')}{{p(x_t) Q(x'; x_t)}} \]

\[ x_{t+1} = x' \text{ with probability } \min(\alpha, 1) \]

\[ x_{t+1} = x_t \text{ otherwise} \]
Metropolis–Hastings algorithm

\[
\alpha = \frac{p(x')Q(x_t; x')}{p(x_t)Q(x'; x_t)}
\]

\[
x_{t+1} = x' \text{ with probability } \min(\alpha, 1)
\]

\[
x_{t+1} = x_t \text{ otherwise}
\]
tempered simulation
• **Differential Evolution algorithm** (Ter Baark 2006)

• **Hybrid MCMC algorithm** – tempered differential evolution

• A typical run has 15-20 free parameters. 256 chains run in parallel. Converge around 4000 iterations.
posterior predictive check

- posterior predictive distribution should agree with observational data
- quantify the probability of having the data given the model is true.
  - define a test statistic
  - compute the reference distribution of the test statistic using posterior
  - define a probability measure (p-value) and compute it

\[
p_B = P[T(D^{\text{rep}}) \geq T(D)|D_c] = \int \int I_{T(D^{\text{rep}}) \geq T(D)} p(D^{\text{rep}}|\Theta) p(\Theta|D_c) dD^{\text{rep}} d\Theta
\]
posterior predictive check

• posterior predictive distribution should agree with observational data

• quantify the probability of having the data given the model is true.
  - define a test statistic
  - compute the reference distribution of the test statistic using posterior
  - define a probability measure (p-value) and compute it
posterior predictive check

- posterior predictive distribution should agree with observational data
- quantify the probability of having the data given the model is true.
  - define a test statistic
  - compute the reference distribution of the test statistic using posterior
  - define a probability measure (p-value) and compute it

\[ \hat{p}_B = \frac{1}{L} \sum_{l=1}^{L} I_{T(D^{rep}_l) \geq T(D)} \]
prediction: stellar mass functions

Li & White 2009

Perez-Gonzalez et al 2008

Marchesini et al 2009

Stark et al 2009
PPC for stellar mass functions
stellar mass functions in terms of Schechter parameters
prediction: cold gas mass function

HIPASS data
cosmic SFR density  cosmic cold gas mass density
\[ \Delta M_{\text{out}} = \alpha_{\text{RH}} \left( \frac{200 \text{km/s}}{V_c} \right)^{\beta_{\text{RH}}} \Delta M_\ast \]

\[ \Delta E_{\text{out}} = \begin{cases} 
\frac{1}{2} \Delta M_{\text{out}} V_{\text{esc}}^2 & \text{ejection;} \\
\frac{5}{4} \Delta M_{\text{out}} V_c^2 & \text{reheating.} 
\end{cases} \]

\[ \Delta E_{\text{FB}} = \epsilon_{FB} \eta_{\text{SN}} E_{\text{SN}} \Delta m_\ast \]

\[ \Delta E_{\text{FB}} \geq \Delta E_{\text{out}} \]
• observations show OFR ~ SFR; outflow rates in the cold component of the wind is of the order of 10% of the star formation rate (Martin et al. 2006).

• recent theory shows outflow mass loading can only be as large as about unity (Dekel & Krumholz 2013).
\[ M_h = 10^{11} M_\odot \]
\[ M_* \approx 0.004 M_h \]
\[ M_{\text{cold}} \approx 10 M_* \]

\[ \frac{M_{\text{outflow}}}{M_{\text{SF}}} \approx \frac{0.17 - 0.004 \times 10 - 0.004}{0.004 \times 2} \approx 16 \]

Papastergis et al. 2012
Conclusions

• Many physical processes affecting galaxy formation are not yet well understood while copious observational data are available to constrain models. The problem is best tackled with the Bayesian inference approach.

• The Bayesian SAM provides an approach to constraining galaxy formation models with observational data in a statistically rigorous way.

• The Bayesian approach can be used to probe the tension between data and model and to help identify missing physics.

• For a given model family, the posterior distribution of model parameters obtained for a given set of data can be used to predict observables that include the inferential uncertainties. Such predictions can be used to assess the power of new observations.

• Outstanding issues/difficulties: slow likelihood evaluation, definition of likelihood function, where to meet theory with observation...
Thank you!