Hierarchical Bayesian Modeling

Making scientific inferences about a population based on many individuals

Angie Wolfgang
NSF Postdoctoral Fellow, Penn State
Astronomical Populations

Once we discover an object, we look for more . . .

Lissauer, Dawson, & Tremaine, 2014

Schawinski et al. 2014

to characterize their properties and understand their origin.
Astronomical Populations

Or we use many (often noisy) observations of a single object to gain insight into its physics.
Hierarchical Modeling is a statistically rigorous way to make scientific inferences about a population (or specific object) based on many individuals (or observations).

Frequentist multi-level modeling techniques exist, but we will discuss the Bayesian approach today.

**Frequentist: variability of sample**
(If \( \_ \) is the true value, what fraction of many hypothetical datasets would be as or more discrepant from \( \_ \) as the observed one?)

**Bayesian: uncertainty of inference**
(What’s the probability that \( \_ \) is the true value given the current data?)
Understanding Bayes

Bayes’ Theorem
(straight out of conditional probability)

\[ p(\theta|x) \propto p(x|\theta) \, p(\theta) \]

posterior  likelihood  prior

\( x = \) data
\( \theta = \) the parameters of a model that can produce the data
\( p() = \) probability density distribution of
\( | = \) “conditional on”, or “given”

\[ p(\theta) = \text{prior probability} \]
(How probable are the possible values of \( \theta \) in nature?)

\[ p(x|\theta) = \text{likelihood, or sampling distribution} \]
(Ties your model to the data probabilistically:
how likely is the data you observed given specific \( \theta \) values?)

\[ p(\theta|x) = \text{posterior probability} \]
(A “new prior” distribution, updated with information contained in the data:
what is the probability of different \( \theta \) values given the data and your model?)
Applying Bayes

\[ p(\theta|x) \propto p(x|\theta) \ p(\theta) \]

posterior  likelihood  prior

Example (1-D): Fitting an SED to photometry

\( x = 17 \) measurements of \( L_\nu \)

Model: Stellar Population Synthesis

\( \theta \) = age of stellar population,
star formation timescale \( \tau \),
dust content \( A_V \),
metallicity,
redshift,
choice of IMF,
choice of dust reddening law

Model can be summarized as \( f(x|\theta) \):
Maps \( \theta \to x \).

But this is NOT \( p(x|\theta) \) because
\( f(x|\theta) \) is not a probability distribution!!
Applying Bayes

\[ p(\theta|\mathbf{x}) \propto p(\mathbf{x}|\theta) \, p(\theta) \]

posterior \quad likelihood \quad prior

Example (1-D): Fitting an SED to photometry

\[ \mathbf{x} = 17 \text{ measurements of } L_{\nu} \]

Model: Stellar Population Synthesis

\( \theta = \) age of stellar population,
    star formation timescale \( \tau \),
    dust content \( A_v \),
    metallicity,
    redshift,
    choice of IMF,
    choice of dust reddening law

Model can be summarized as \( f(x|\theta) \):
    Maps \( \theta \rightarrow \mathbf{x} \).

But this is NOT \( p(x|\theta) \) because \( f(x|\theta) \) is not a probability distribution!!

If use \( \chi^2 \) for fitting, then you are implicitly assuming that:

\[ p(x_i|\theta) = \frac{1}{\sqrt{2\pi}\sigma_i^2} e^{-\frac{(x_i-\mu)^2}{2\sigma_i^2}} \]

where \( \mu = f(x_i|\theta) \)
and \( \sigma_i = \) “statistical measurement error”
i.e. you are assuming “Gaussian noise”
(if you could redo a specific \( x_i \)
the same way many times, you’d find:)

![Graphical representation of the distribution of \( x \) with mean \( \mu \) and standard deviation \( \sigma \).]
Applying Bayes

\[ p(\theta|x) \sim p(x|\theta) \ p(\theta) \]

\( p(x|\theta) \) = posterior  \( p(x|\theta) \) = likelihood  \( p(\theta) \) = prior

Example (2-D): Fitting a PSF to an image

\( x \) = matrix of pixel brightnesses

\( \theta = \mu, \sigma \) of Gaussian (location, FWHM of PSF)

\( f(x|\theta) = 2\text{-D Gaussian} \)

\[ p(x|\theta) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right) \]

where \( \mu = f(x|\theta) \) and \( \Sigma = \text{noise} \) (possibly spatially correlated)

Both likelihood and model are Gaussian!!

Raw Data

Gaussian Fit Function

Localized Data Point

(c)
Applying Bayes

\[ p(\theta|x) \propto p(x|\theta) \ p(\theta) \]

posterior likelihood prior

Example (1-D): Fitting an SED to photometry

\( x = 17 \) measurements of \( L_\nu \)

Model: Stellar Population Synthesis

\( \theta = \) age of stellar population,
star formation timescale \( \tau \),
dust content \( A_V \),
metallicity,
redshift,
choice of IMF,
choice of dust reddening law

Model can be summarized as \( f(x|\theta) \):
Maps \( \theta \rightarrow x \).

But this is NOT \( p(x|\theta) \) because \( f(x|\theta) \) is not a probability distribution!!

Ok, now we know of one way to write \( p(x|\theta) \).

What about \( p(\theta) \)?

1) If we have a previous measurement/inference of that object’s metallicity, redshift, etc., use it \textit{with its error bars} as \( p(\theta) \).
(Usually “measured” via \( \chi^2 \), so \( p(\theta) \) is Gaussian with \( \mu = \) measurement and \( \sigma = \) error. BUT full posteriors from previous analysis is better.)

2) Choose wide, uninformative distributions for all the parameters we don’t know well.

3) Use distributions in nature from previous observations of similar objects.
Going Hierarchical

Option #3 for $p(\theta)$:
Use distributions in nature from previous observations of similar objects.

Histories of population properties, when normalized, can be interpreted as probability distributions for individual parameters:

$$p(\theta) = \frac{n(\theta|\alpha)}{\int n(\theta|\alpha)d\theta} = p(\theta|\alpha)$$

where $n(\theta|\alpha)$ is the function with parameters $\alpha$ that was fit to the histogram (or even the histogram itself, if you want to deal with a piecewise function!)

For example, redshift was part of the $\theta$ for SED fitting. One could use the red lines (parametric form below) as

$$p(z) = p(z|\alpha) = \frac{n(z|\alpha)}{\int n(z|\alpha)dz}$$

with $n(z|\alpha) = \frac{z^a + z^{ab}}{z^b + c}$ and $\alpha = \{a,b,c\}$.

But BE CAREFUL of detection bias, selection effects, upper limits, etc.!!!!!!
Going Hierarchical

Option #3 for \( p(\theta) \):
Use distributions in nature from previous observations of similar objects.

Histograms of population properties, when normalized, can be interpreted as probability distributions for individual parameters:

\[
p(\theta) = \frac{n(\theta|\alpha)}{\int n(\theta|\alpha)d\theta} = p(\theta|\alpha)
\]

where \( n(\theta|\alpha) \) is the function with parameters \( \alpha \) that was fit to the histogram (or even the histogram itself, if you want to deal with a piecewise function!)

Abstracting again ….

\[
p(\theta|x) \propto p(x|\theta) \, p(\theta)
\]

\[
\text{posterior} \quad \text{likelihood} \quad \text{prior}
\]

\[
p(\theta|x) \propto p(x|\theta) \, p(\theta|\alpha)
\]

\[
\text{posterior} \quad \text{likelihood} \quad \text{prior}
\]

(April there!!)

Population helps make inference on individual …
Going Hierarchical

... but what if we want to use the individuals to infer things (the $\alpha$’s) about the population?

i.e., $p(\theta|\alpha)$ contains some interesting physics and getting values for $\alpha$ given the data can help us understand it.

$$p(\theta|x) \propto p(x|\theta) \, p(\theta|\alpha)$$  

posterior likelihood prior

$$p(\alpha,\theta|x) \propto p(x|\theta,\alpha) \, p(\theta|\alpha) \, p(\alpha)$$  

posterior likelihood prior

If you truly don’t care about the parameters for the individual objects, then you can marginalize over them:

$$p(\alpha|x) \propto \left[ \int p(x|\theta,\alpha) \, p(\theta|\alpha) \, d\theta \right] \, p(\alpha) = p(x|\alpha) \, p(\alpha)$$  

posterior likelihood prior
Graphically:

“Regular” Bayes:

\[ p(\theta|x) \propto p(x|\theta) \, p(\theta) \]
posterior likelihood prior

Hierarchical Bayes:

\[ p(\alpha,\theta|x) \propto p(x|\theta,\alpha) \, p(\theta|\alpha) \, p(\alpha) \]
posterior likelihood prior
Graphically:

"Regular" Bayes:

\[ p(\theta|x) \propto p(x|\theta) \, p(\theta) \]

posterior  likelihood  prior

Hierarchical Bayes:

\[ p(\alpha,\theta|x) \propto p(x|\theta,\alpha) \, p(\theta|\alpha) \, p(\alpha) \]

posterior  likelihood  prior

Even for an individual object, connection between parameters and observables can involve several layers. (Example: measuring orbital parameters of stars in a binary system)

Parameters

Latent Variables

Observables

Conditional independence between individuals:

Population Parameters

Individual Parameters

Observables
HBM in Action: Model

Exoplanet compositions: Wolfgang & Lopez, 2015

\[ p(\theta, \alpha|X) = p(\{R_{pl,i}, M_{core,i}, f_{env,i}\}, \{\alpha, \mu, \sigma, \gamma\}|\delta_i, \sigma_{\delta,i}, F_i) \propto \]

\[ \prod_{i=1}^{N} \left\{ p(\delta_i|\sigma_{\delta,i}, R_{pl,i}, R_{*i}, M_{core,i}, f_{env,i}, F_i, \alpha, \mu, \sigma, \gamma) \right\} \]

\[ \times \prod_{i=1}^{N} \left\{ p(R_{*i})p(M_{pl,i}|\alpha)p(f_{env,i}|\mu, \sigma) \right\} p(\alpha)p(\mu)p(\sigma)p(\gamma) \]

- compositions of individual super-Earths (fraction of mass in a gaseous envelope: \( f_{env} \))

- the distribution of this composition parameter over the Kepler population (\( \mu, \sigma \)).
HBM in Action: Results

Exoplanet compositions: Wolfgang & Lopez, 2015

Posterior on population parameters:

Marginal composition distribution:

Width of distribution had not been previously characterized.
HBM in Action: Results

Exoplanet compositions: Wolfgang & Lopez, 2015

Posteriors on composition parameter $f_{\text{env}}$ for individual planets:
A Note About Shrinkage

Hierarchical models “pool” the information in the individual data …

... which shrinks individual estimates toward the population mean and lowers overall RMS error. (A key feature of any multi-level modeling!)
A Note About Shrinkage

Shrinkage in action:

uncertainty in $x_1$ when analyzed in hierarchical model

uncertainty in $x_1$ when analyzed by itself

mean of distribution of x's

Gray = data
Red = posteriors

Wolfgang, Rogers, & Ford, 2016

Wolfgang, Rogers, & Ford, 2016

Mean (M_{Earth})

Radius (R_{Earth})
Practical Considerations

1) Pay attention to the structure of your model!!
   - Did you capture the important dependencies and correlations?
   - Did you balance realism with a small number of population-level parameters?

2) Evaluating your model with the data (performing hierarchical MCMC):
   - JAGS ([http://mcmc-jags.sourceforge.net](http://mcmc-jags.sourceforge.net); can use stand-alone binary or interface with R)
   - STAN ([http://mc-stan.org](http://mc-stan.org); interfaces with R, Python, Julia, MATLAB)
   - Or write your own hierarchical MCMC code

3) Spend some time testing the robustness of your model: if you generate hypothetical datasets using your HBM and then run the MCMC on those datasets, how close do the inferences lie to the “truth”??
More About STAN

Based on the Hamiltonian Monte Carlo (HMC) algorithm (More at Betancourt 2017, arXiv:1701.02434)

What is HMC? How does it compare to Metropolis-Hastings?

Consider the following probability density* \( p(x|\theta) = \pi(q) \)

You want to sample the high probability regions (red) efficiently:

A random walk (Metropolis-Hastings) won’t follow the curve as easily, and so won’t be efficient:

*In high dimensions, the relevant quantity is actually the probability density scaled by the volume of parameter space that we have to probe: \( \pi(q) \, dq \). This is called the “typical set”. The typical set in higher dimensions is much more likely to look like the above than the probability density \( \pi(q) \) by itself.
Hamiltonian Monte Carlo

So how do we get the samples to follow the curve? As physicists, we know how to compute orbital trajectories …

Use the Hamiltonian to track particle locations over time!

\[
\frac{dq}{dt} = + \frac{\partial H}{\partial p} \quad \text{time evolution of a particle's position}
\]

\[
\frac{dp}{dt} = - \frac{\partial H}{\partial q} \quad \text{time evolution of a particle's momentum}
\]

In physics, we know what \( H \) is supposed to be:

\[
\mathcal{H} = K + V \quad , \quad K = \frac{p^2}{2m} \quad , \quad V = V(q)
\]

What is \( H \) in probability space?
Hamiltonian Monte Carlo

Introduce auxiliary “momentum” parameters to the likelihood $\pi(q)$:

$$q_n \rightarrow (q_n, p_n)$$

Assert an easily useable form for the new prob. density $\pi(q,p)$:

$$\pi(q,p) = e^{-H(q,p)}$$

Then complete the analogy:

$$H(q,p) = -\log \pi(p \mid q) - \log \pi(q)$$

$$\equiv K(p,q) + V(q).$$

In practice, this involves computing gradients of the likelihood:

\[
\begin{align*}
\frac{dq}{dt} &= + \frac{\partial H}{\partial p} = \frac{\partial K}{\partial p} \\
\frac{dp}{dt} &= - \frac{\partial H}{\partial q} = - \frac{\partial K}{\partial q} - \frac{\partial V}{\partial q}
\end{align*}
\]

\(\frac{\partial V}{\partial q}\) is the gradient of $\pi(q)$

Also, if we choose $\pi(p \mid q) = \mathcal{N}(p \mid 0, \Sigma(q))$ then:

$$K(q,p) = \frac{1}{2}p^T \cdot \Sigma^{-1}(q) \cdot p + \frac{1}{2} \log |\Sigma(q)| + \text{const.}$$
So what does this look like?

Let's say our likelihood $\pi(q)$ looks like this:

Let's watch as the sampler traces probability space:
https://chi-feng.github.io/mcmc-demo/app.html#HamiltonianMC,banana

Another website with animations: http://elevanth.org/blog/2017/11/28/build-a-better-markov-chain/
In Sum, Why HBM?

- Obtain *simultaneous posteriors* on individual and population parameters: self-consistent constraints on the physics

- Readily *quantify uncertainty* in those parameters

- Naturally deals with *large measurement uncertainties and upper limits* (censoring)

- Similarly, can *account for selection effects* *within* the model, simultaneously with the inference

- Enables *direct, probabilistic relationships* between theory and observations

- Framework for *model comparison*
Further Reading

**Introductory/General:**

DeGroot & Schervish, *Probability and Statistics*  
(Solid fundamentals)

Gelman, Carlin, Stern, & Rubin, *Bayesian Data Analysis*  
(In-depth; advanced topics)

(Few-page intro/overview of multi-level modeling in astronomy)

B.C. Kelly 2007  
(HBM for linear regression, also applied to quasars)

Betancourt 2017 (arXiv:1701.02434)  
(Intro to Hamiltonian Monte Carlo)

**Some applications:**

Loredo & Wasserman, 1998  
(Multi-level model for luminosity distribution of gamma ray bursts)

Mandel et al. 2009  
(HBM for Supernovae)

Hogg et al. 2010  
(HBM with importance sampling for exoplanet eccentricities)

Andreon & Hurn, 2010  
(HBM for galaxy clusters)

Martinez 2015  
(HBM for Milky Way satellites)

Wolfgang, Rogers, & Ford 2016  
(HBM for exoplanet mass-radius relationship)
On to the Lab . . .

Hierarchical Bayesian Modeling with STAN:
Inferring the Exoplanet Mass Distribution from Radial Velocities