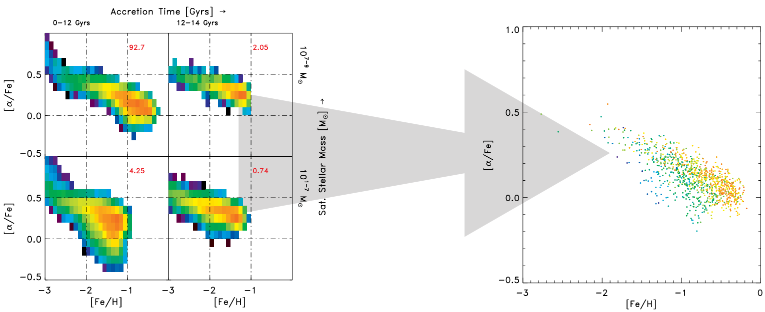
A generative finite mixture model

Theoretical halo distributions

Observations



$$\left[\frac{Fe}{H}, \frac{\alpha}{Fe}\right]_{i=1}^{N} \text{i.i.d} \sim f(x,y) = \sum_{j=1}^{m} \pi_j f_j(x,y)$$

A generative finite mixture model

define x

$$\left[\frac{Fe}{H}, \frac{\alpha}{Fe}\right]_{i=1}^{N} \text{i.i.d} \sim f(x,y) = \sum_{j=1}^{m} \pi_j f_j(x,y)$$

where the mixing proportions, $\boldsymbol{\pi}$, give the formation history of the halo

$$\sum_{j=1}^{m} \pi_j = 1, \quad \pi_j \geq 0, \quad j = 1, \ldots, m$$

Estimating the mixing proportions $\boldsymbol{\pi}$

To estimate the mixing proportions, we can use a maximum likelihood approach

$$\hat{\boldsymbol{\pi}}_{\text{MLE}} = \underset{\boldsymbol{\pi}}{\operatorname{argmax}}\, L(\boldsymbol{\pi})$$

where $$L(\boldsymbol{\pi}) = \sum_{i=1}^{n} \log\Big(\sum_{j=1}^{m} \pi_j f_j(x_i, y_i)\Big)$$

Unfortunately the standard MLE procedure for estimating $\boldsymbol{\pi}$ is intractable with this likelihood.

Expectation Maximization (EM) algorithm to the rescue!

Adding the latent variable $\mathbf{z}$

Suppose we knew which mixture component f_j each observation came from:

$$
\begin{aligned}
z_{ij} = 1 &\quad \text{if } (x_i, y_i) \sim f_j \\
0 &\quad \text{otherwise}
\end{aligned}
$$

The log likelihood can then be expressed as

$$L(\boldsymbol{\pi}) = \sum_{i=1}^{n} \sum_{j=1}^{m} z_{ij} \log \left\{ \pi_j f_j(x_i, y_i) \right\}$$

The addition of the latent variable $\mathbf{z}$ actually makes things easier because it is easily differentiable in $\boldsymbol{\pi}$.

Finding $\hat{\boldsymbol{\pi}}$ using expectation maximization

but we dont know z

$$\hat{\boldsymbol{\pi}}^{(t)} = \underset{\boldsymbol{\pi}}{\operatorname{argmax}} \, \mathbb{E}_{\mathbf{z}|\mathbf{x},\mathbf{y},\hat{\boldsymbol{\pi}}^{(t-1)}} \Big[L(\boldsymbol{\pi}) \Big]$$

At each time t, the estimate for $\boldsymbol{\pi}$ is the argmax of the expected log likelihood, given the data and the estimate of $\boldsymbol{\pi}$ from the previous time.

- First we find the expected value of $L(\boldsymbol{\pi})$ using the current expected values of the latent variable $\mathbf{z}$
- $\hat{\boldsymbol{\pi}}^{(t)}$ is the $\operatorname{argmax}_{\boldsymbol{\pi}}$ of this expectation, which is simple to compute
- We repeat until $L(\boldsymbol{\pi})$ stabilizes

Find the expected value of $L(\boldsymbol{\pi})$ using the current estimate of the latent variable

The expected value of $L(\boldsymbol{\pi})$, with respect to the conditional distribution of $\mathbf{z}$, given observed data and $\boldsymbol{\pi}^{(t-1)}$ is

$$\mathbb{E}_{\boldsymbol{\pi}}\Big[L(\boldsymbol{\pi})\Big|\mathbf{x},\mathbf{y}\Big] = \sum_{i=1}^{n}\sum_{j=1}^{m}\mathbb{E}_{\boldsymbol{\pi}}\left[z_{ij}|x_i,y_i\right]\left\{\log f_j(x_i,y_i) + \log \pi_j\right\}$$

Since z_{ij} is an indicator, its expected value is simply the probability that data point i comes from model j

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\pi}}\Big[z_{ij}|x_i,y_i\Big] &= \Pr_{\boldsymbol{\pi}}(z_{ij}|x_i,y_i) \\
&= \frac{p(x_i,y_i|z_{ij}=1)p(z_{ij}=1)}{p(x_i,y_i)} \\
&= \frac{\pi_j f_j(x_i,y_i)}{\sum_{j=1}^{m}\pi_j f_j(x_i,y_i)}
\end{aligned}$$

Find the argmax$_{\pi}$ of this expectation

Now that we have the expected value of $L(\boldsymbol{\pi})$ with respect to the conditional distribution of $\mathbf{z}$, we need only evaluate

$$\hat{\boldsymbol{\pi}}^{(t)} = \underset{\boldsymbol{\pi}}{\operatorname{argmax}}\, \mathbb{E}\left[L(\boldsymbol{\pi})\middle|\mathbf{x},\mathbf{y},\hat{\boldsymbol{\pi}}^{(t-1)}\right]$$

Let

$$\hat{\mathbf{w}}^{(t)} = \mathbb{E}_{\boldsymbol{\pi}}\left[z_{ij}\middle|x_i,y_i\right]$$

Accounting for the $m-1$ free parameters of $\boldsymbol{\pi}$, differentiation proceeds, for $k=1,\ldots,m-1$, as:

$$\frac{\partial}{\partial\pi_k}\mathbb{E}\left[L(\boldsymbol{\pi})\middle|\mathbf{x},\mathbf{y}\right] = \sum_{i=1}^{n}\left\{w_{ik}^{(t-1)}\frac{1}{\pi_k} - w_{im}^{(t-1)}\frac{1}{1-\pi_1-\ldots-\pi_{m-1}}\right\}$$

$$\frac{1}{\pi_k}\sum_{i=1}^{n} w_{ik}^{(t-1)} = \frac{1}{1-\pi_1-\ldots-\pi_{m-1}}\sum_{i=1}^{n} w_{im}^{(t-1)}$$

Find the argmax$_\pi$ of this expectation

Consequently

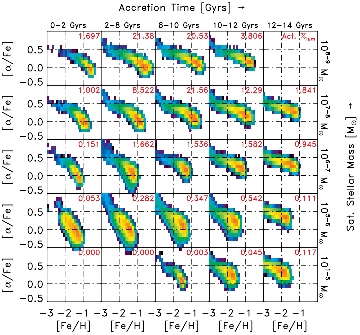
$$\hat{\pi}_k^{(t)} = \begin{cases} \frac{\sum_{i=1}^{n} w_{ij}^{(t-1)}}{n} & k = 1, \ldots, m-1 \\ 1 - \pi_1 - \ldots - \pi_{m-1} & k = m \end{cases}$$

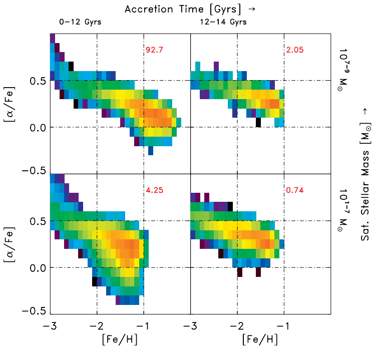
Where

$$w_{ij}^{(t+1)} = \begin{cases} \dfrac{\pi_j^{(t)} f_j(x_i, y_i)}{\sum\limits_{k=1}^{m} \pi_k^{(t)} f_k(x_i, y_i)} & j = 1, \ldots, m-1 \\ 1 - w_{i1} - \ldots - w_{i,m-1} & j = m \end{cases}$$

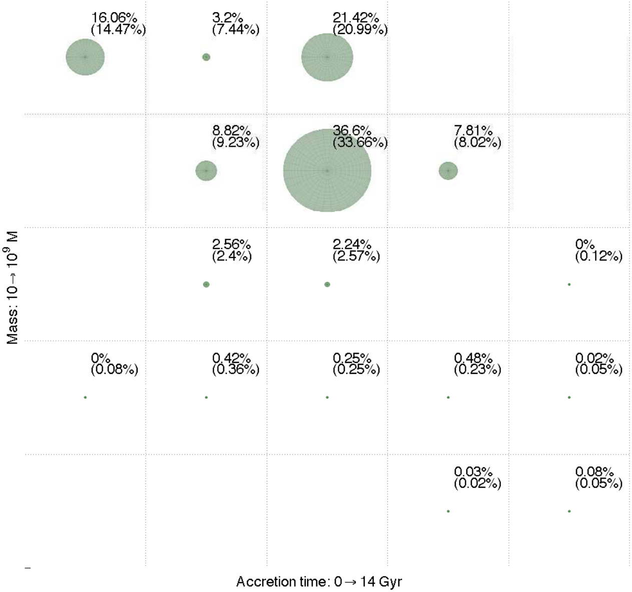
Simulations

- Generated observations from one halo and multiple halos
- Used a 5x5 grid ($m = 25$), and several 2x2 grids ($m = 4$)
  - 5x5 grid did not work for some halo realizations
  - 2x2 grid reliably converged on the correct mixing proportions

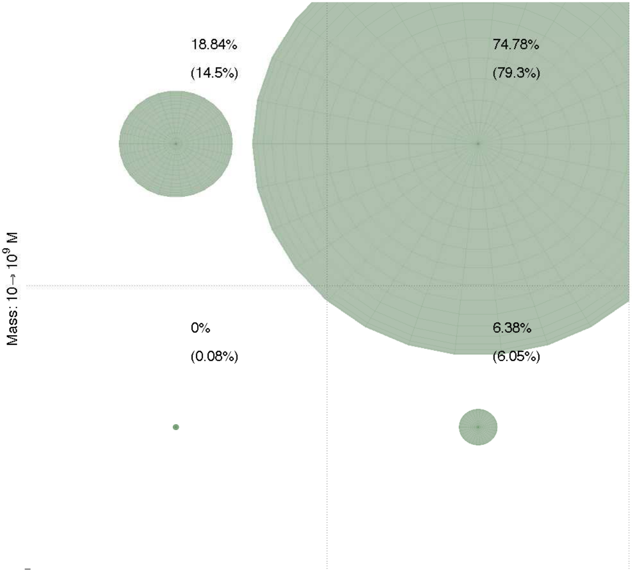




EM formation history for m=25

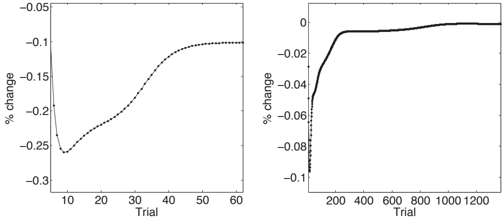


EM formation history for $m=4$

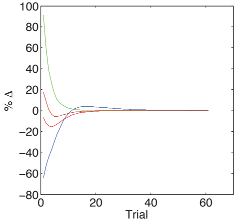


Simulation overview

Log likelihood % change

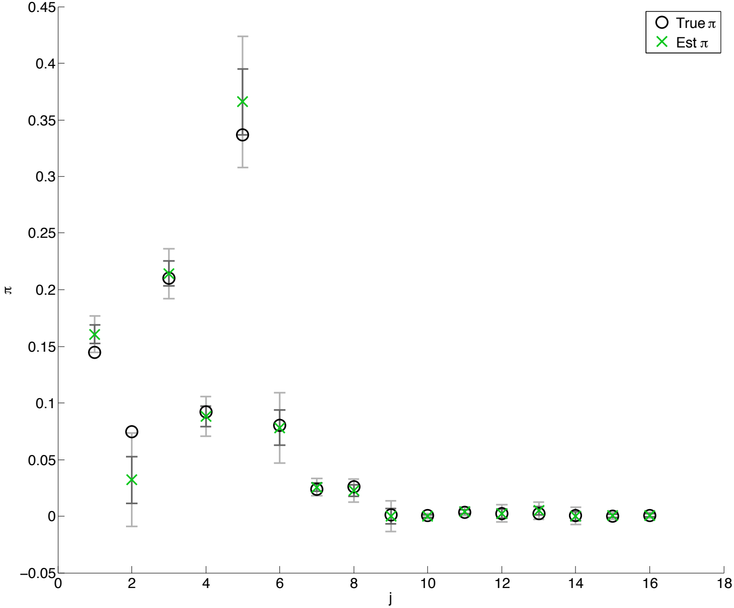


π % change

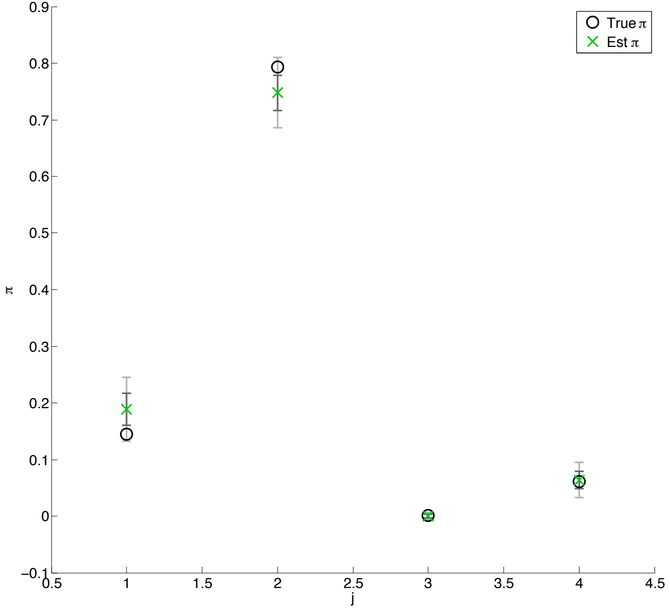


- Works with as few as 1,000 observations
- Insensitive to initialization of π
- Always converges
- Large weights identified after 10 iterations
- $L(\pi)$ stops changing appreciably after 60 (m=4) or 600 (m=25) iterations

Confidence intervals from observed Fisher Information

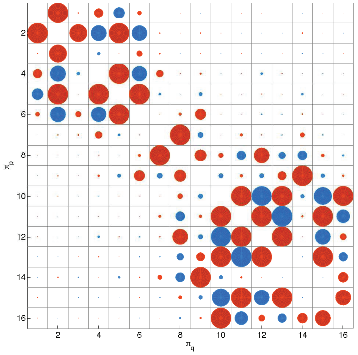


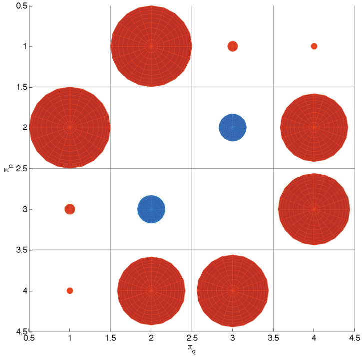
Confidence intervals from observed Fisher Information



Correlation between π

$$I(\pi'|\mathbf{x},\mathbf{y}) = -\frac{\partial^2 L(\pi')}{\partial \pi' \partial \pi'^T}$$





Conclusion

Worked

- 2x2
- EM
- 5x5 in a few cases
- M-of-n bootstrapped errors

Did not work

- 5x5
- Parametric bootstrapped errors

Future improvements

- Non-arbitrary gridding
- Smoothing of f_j