

METALLICITY

OCTOBER 24, 2010

1. MIXTURE MODEL

For each observation of $(L, \frac{\alpha}{\text{Fe}}, \frac{\text{Fe}}{\text{H}})$, let $\{(x_i, y_i)\}_{i=1}^n$ represent observed metallicities of stars drawn from one of m known model densities. We model the density of observations using the mixture model

$$f(x, y) = \sum_{j=1}^m \pi_j f_j(x, y) \quad \sum_{j=1}^m \pi_j = 1 \quad \pi_j \geq 0, j = 1, \dots, m$$

With a complete likelihood of

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

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

Evaluation of $\partial l(\boldsymbol{\pi}) / \partial \pi$ can be avoided by adding a latent indicator, z , to the observed data (x, y) , representing the model group from which that observation was generated. Let G_j be the j^{th} model group, and let

$$z_{ij} = \mathbf{1}\{(x_i, y_i) \mapsto G_j\}$$

The incomplete data likelihood is defined over the complete data $\{(x_i, y_i, \mathbf{z}_i)\}_{i=1}^n$ as

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

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

2. EM

First we find the expected value of $l(\boldsymbol{\pi})$ conditional on the distribution of \mathbf{z} . Since z_{ij} is an indicator function, its expected value is equal to the probability that data point i comes from model j .

2.1. Expected value of $l(\boldsymbol{\pi})$. The expected value of $l(\boldsymbol{\pi})$ is

$$\begin{aligned} E_{\boldsymbol{\pi}}[l(\boldsymbol{\pi})|\mathbf{x}, \mathbf{y}] &= \sum_{i=1}^n E(\boldsymbol{\pi}|x_i, y_i) l(\boldsymbol{\pi}|x_i, y_i) \\ &= \sum_{i=1}^n \sum_{j=1}^m E_{\boldsymbol{\pi}}[z_{ij}|x_i, y_i] \{ \log f_j(x_i, y_i) + \log \pi_j \} \end{aligned}$$

where

$$E_{\boldsymbol{\pi}}[z_{ij}|x_i, y_i] = \text{Probability}\left((x_i, y_i) \mapsto G_j | x_i, y_i\right)$$

2.2. Expected value of $\pi|\mathbf{x}, \mathbf{y}$. The expected value of z_{ij} is the same as the expected value of π_j , given the data. This can be specified as

$$E_{\boldsymbol{\pi}}[z_j|x_i, y_i] = P(z_j|x_i, y_i) = \frac{P(x_i, y_i|z_j = 1)P(z_j = 1)}{P(x_i, y_i)}$$

with constituent parts:

$$P(x_i, y_i|z_j = 1) = f_j(x_i, y_i) \quad P(x_i, y_i|z_j) = \prod_{j=1}^m f_j^{z_j} \quad P(z_j) = \prod_{j=1}^m \pi_j^{z_j}$$

thus

$$E_{\boldsymbol{\pi}}[z_j|x_i, y_i] = \frac{\pi_j f_j(x_i, y_i)}{\sum_{j=1}^m \pi_j f_j(x_i, y_i)}$$

Defining $w_{ij}^{(t)}$ as the expected value of z_{ij} at the t^{th} step, and $\pi_j^{(t)}$ as the MLE of π_j at the t^{th} step, yeilds

$$w_{ij}^{(t+1)} = \frac{\pi_j^{(t)} f_j(x_i, y_i)}{\sum_{k=1}^m \pi_k^{(t)} f_k(x_i, y_i)}$$

2.3. Solving for π .

$$\begin{aligned} 0 &= \frac{\partial}{\partial \pi_k} E_{\pi} [l(\pi) | \mathbf{x}, \mathbf{y}] \\ &= \sum_{i=1}^n \left\{ w_{ij}^{(0)} \frac{1}{\pi_k} - w_{im}^{(0)} \frac{1}{1 - \pi_1 - \dots - \pi_{m-1}} \right\}, k = 1, \dots, m-1 \end{aligned}$$

Therefore

$$\begin{aligned} \frac{1}{\pi_1} \sum_{i=1}^n w_{i1}^{(0)} &= \dots = \frac{1}{\pi_{m-1}} \sum_{i=1}^n w_{i,m-1}^{(0)} = c \\ \hat{\pi}_k &= \frac{\sum_{i=1}^n w_{ik}^{(0)}}{c} \\ \pi_j^{(1)} &= \frac{\sum_{i=1}^n w_{ij}^{(0)}}{n} \end{aligned}$$

And in general,

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^n w_{ij}^{(t)}}{n}$$

3. OBSERVED INFORMATION

As $\pi_m = 1 - \sum_{j=1}^m \pi_j$ there are only g free parameters. Thus let $g = m - 1$, and $\boldsymbol{\pi}' = \boldsymbol{\pi}[1, \dots, g]$. The observed information matrix, $I(\boldsymbol{\pi}'|\mathbf{x}, \mathbf{y})$ is given by the $g \times g$ negative hessian:

$$\begin{aligned}
 l(\boldsymbol{\pi}') &= \sum_{i=1}^n \log \left(\sum_{j=1}^g \pi_j f_j \right) \\
 \frac{\partial l(\boldsymbol{\pi}')}{\partial \pi_k} &= \sum_{i=1}^n \frac{f_k}{\sum_{j=1}^g \pi_j f_j} \\
 \frac{\partial^2 l(\boldsymbol{\pi}')}{\partial \pi_k \partial \pi_r} &= - \sum_{i=1}^n \frac{f_k}{(\sum_{j=1}^g \pi_j f_j)^2 f_r} \\
 I(\boldsymbol{\pi}|\mathbf{x}, \mathbf{y}) &= - \frac{\partial^2 l(\boldsymbol{\pi}')}{\partial \boldsymbol{\pi}' \partial \boldsymbol{\pi}'^T} = - \begin{bmatrix} \frac{\partial^2 l(\boldsymbol{\pi}')}{\partial \pi_1^2} & \frac{\partial^2 l(\boldsymbol{\pi}')}{\partial \pi_1 \partial \pi_2} & \cdots & \frac{\partial^2 l(\boldsymbol{\pi}')}{\partial \pi_1 \partial \pi_g} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 l(\boldsymbol{\pi}')}{\partial \pi_g \partial \pi_1} & \frac{\partial^2 l(\boldsymbol{\pi}')}{\partial \pi_g \partial \pi_2} & \cdots & \frac{\partial^2 l(\boldsymbol{\pi}')}{\partial \pi_g^2} \end{bmatrix} \\
 &= \begin{bmatrix} \sum_{i=1}^n \frac{1}{(\sum_{j=1}^g \pi_j f_j)^2} & \sum_{i=1}^n \frac{f_1}{(\sum_{j=1}^g \pi_j f_j)^2 f_2} & \cdots & \sum_{i=1}^n \frac{f_1}{(\sum_{j=1}^g \pi_j f_j)^2 f_g} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^n \frac{f_g}{(\sum_{j=1}^g \pi_j f_j)^2 f_1} & \sum_{i=1}^n \frac{f_g}{(\sum_{j=1}^g \pi_j f_j)^2 f_2} & \cdots & \sum_{i=1}^n \frac{1}{(\sum_{j=1}^g \pi_j f_j)^2} \end{bmatrix}
 \end{aligned}$$

*Excluding i where $(\sum_{j=1}^g \pi_j f_j)^2 f_i = 0$ to avoid dividing by zero.

3.1. Variance and Covariance of $\boldsymbol{\pi}$. The asymptotic covariance matrix of $\boldsymbol{\pi}$ can be approximated by the inverse of the observed information matrix, $I^{-1}(\hat{\boldsymbol{\pi}}|\mathbf{x}, \mathbf{y})$, yielding

$$\text{Cov}(\pi_p, \pi_q) = \begin{cases} I^{-1}(\hat{\boldsymbol{\pi}}) & p, q < m \\ \sum_{j=1}^g \text{Cov}(\pi_j, \pi_q) & p = m, q < m \\ \sum_{j=1}^g \sum_{k=1}^g \text{Cov}(\pi_j, \pi_k) & p, q = m \end{cases}$$

$$\text{Var}(\pi_j) = \sigma_j^2 = \left\{ |\text{Cov}(\boldsymbol{\pi})| \right\}_{jj}$$

* There are many negative covariances; this is ok I think.

4. LIKELIHOOD RATIO TEST

Given certain conditions

$$H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_{\text{true}}$$

$$H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}_{\text{true}}$$

$$\Lambda = -2 \log \frac{\sup_{\boldsymbol{\pi}=\boldsymbol{\pi}_{\text{true}}} l(\boldsymbol{\pi})}{\sup_{\boldsymbol{\pi}} l(\boldsymbol{\pi})} = -2 \{l(\boldsymbol{\pi}_{\text{true}}) - l(\hat{\boldsymbol{\pi}})\} \sim \chi_{m-1}^2$$

$$\Lambda_{\text{Halo } 3} = 25.025 \sim \chi_{15}^2$$

$$\text{p-value } 10\text{k} = 4.961\%$$

$$\text{p-value } 30\text{k} = 1.3 \times 10^{-5}\%$$

$$\text{p-value } 50\text{k} = 1.4 \times 10^{-12}\%$$

Thus we accept H_0 when requiring 95% or less confidence; there is only a 4.961% chance we would see a value this extreme or more given H_0 is true. This holds for 400 to 1600 EM iterations.

5. OTHER

5.1. **Convergence.** $l(\pi)$ from 1.235 to 1.236 over past 25 runs is the slowest rate of convergence.

5.2. **To do.** Bias, consistency, efficiency, method of moments, σ as a function of n . Bootstrapping, $n+1$ bootstrapping as a check that n is large enough for fisher info.

MAP 2.8.5.

Alt T 2.7.3.

basis of test 2.7.1.

deterministic annealing em algo 2.12.4.

stochastic em 2.13.

conditional bootstrap 2.15.5.

alt approach to fisher info: 2.16.1: using wald and reduced parameter holding out.

full bootstrap 2.16.2

2.18.2 modified lrt

2.18.1 outliers

2.19 partial classification

3.5

3.8.1

spectral decomposition: 3.12.1

k means classification for auto binning?

4: mcmc for mixture

5.5.2: exactly what we did–nonnormal

5.5.4: mulitcycle

5.13 hierarchical mixtures of experts

6.13 skewness of lrt and m

6.3.2

6.5

6.6 bootstrapping lrts

6.7 pval boots

6.7.2 double boots

6.8 bias corr for log like, aic, boots info, cv info

6.9 bayeseian info

6.10 classification based info

9 fitting mix to bin data

11 mix analys of directional data

13 hidden markov models