METALLICITY

OCTOBER 26, 2010

1. Mixture model

Given n observed $(\frac{\alpha}{\text{Fe}}, \frac{\text{Fe}}{\text{H}})$ metallicities as $\{(x_i, y_i)\}_{i=1}^n$, or as $(\mathbf{x}, \mathbf{y})$, each of which is drawn from one of m known model densities. We model the density of observations using the mixture model

$$f(x, y) = \sum_{j=1}^{m} \pi_j f_j(x, y) \tag{1}$$

where

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

From the summation constraint, $\boldsymbol{\pi}$ has $m - 1$ free parameters:

$$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{m-1}, 1 - \pi_1 - \ldots - \pi_{m-1})$$

Thus the likelihood of (1) is

$$\begin{aligned}
L(\boldsymbol{\pi}) &= \prod_{i=1}^{n} f(x_i, y_i) \\
&= \prod_{i=1}^{n} \Big\{ \sum_{j=1}^{m} \pi_j f_j(x_i, y_i) \Big\} \\
\log L(\boldsymbol{\pi}) &= \sum_{i=1}^{n} \log \Big(\sum_{j=1}^{m} \pi_j f_j(x_i, y_i) \Big)
\end{aligned}$$

Maximizing $\log L(\boldsymbol{\pi})$ with respect to $\boldsymbol{\pi}$ will yield $\hat{\boldsymbol{\pi}}_{\text{MLE}}$, but this arduous task can be avoided by adding a latent indicator, z, to the observed data $(\mathbf{x}, \mathbf{y})$, representing the model group from which that observation was generated. Let G_j be the j^{th} model group, and let

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

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

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

$$l(\boldsymbol{\pi}) = \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}\log\big\{\pi_j f_j(x_i, y_i)\big\} \tag{2}$$

2. Expectation Maximization

One way to estimate $\boldsymbol{\pi}$ is to use a maximum likelihood estimate, $\hat{\boldsymbol{\pi}}$, computed using expectation maximization. Starting from an initial set of guesses, $\boldsymbol{\pi}^{(0)}$, we iteratively find the expected value of the likelihood, (2), conditional on the data, and then find the $\mathrm{argmax}_{\boldsymbol{\pi}}$ of this expectation. The maximizing value the t^{th} iteration, $\hat{\boldsymbol{\pi}}^{(t)}$, is then used as the starting value for the next run, and we continue until the likelihood changes by less than 10^{-3} over twenty five iterations.

2.1. Expectation step.

Expectation step. First we find the expected value of the log likelihood, (2), conditional on the data. Note that since z_{ij} is an indicator function, its expected value is equal to the probability that data point i comes from model j.

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

Since we're ultimately maximizing, the non-constant component is of primary interest, and can be analytically specified by applying Bayes' rule:

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\pi}}\Big[z_{ij}|x_i, y_i\Big] &= \text{Probability}\Big((x_i, y_i) \mapsto G_j\big|x_i, y_i\Big) \\
&= \mathrm{Pr}_{\boldsymbol{\pi}}(z_{ij}|x_i, y_i) \\
&= \frac{p(x_i, y_i|z_{ij} = 1)p(z_{ij} = 1)}{p(x_i, y_i)}
\end{aligned}$$

Thus the expected value of the indicator variable, z_{ij}, given the data and the parameters, $\boldsymbol{\pi}$, of the data's distribution defined by (1) is

$$\mathrm{E}_{\boldsymbol{\pi}}\Big[z_{ij}|x_i, y_i\Big] = \frac{\pi_j f_j(x_i, y_i)}{\sum_{j=1}^m \pi_j f_j(x_i, y_i)} \tag{3}$$

To iteratively evaluate this expectation, we let $w_{ij}^{(t)}$ be (3) at the t^{th} step:

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

Since $\boldsymbol{\pi}$ is not defined for the first evaluation, we use a random initialization to generate $\mathbf{w}_j^{(0)}$. Convergence is not sensitive to the choice of values in this case, but may be if the likelihood is riddled with local maxima.

2.2. Maximizing with respect to $\boldsymbol{\pi}$.

$$\begin{aligned} 0 &= \frac{\partial}{\partial \pi_k} \mathrm{E}_{\boldsymbol{\pi}}\Big[l(\boldsymbol{\pi})\big|\mathbf{x}, \mathbf{y}\Big] \\ &= \sum_{i=1}^n \Big\{ w_{ij}^{(0)} \frac{1}{\pi_k} - w_{im}^{(0)} \frac{1}{1 - \pi_1 - \ldots - \pi_{m-1}} \Big\}, k = 1, \dots, m-1 \end{aligned}$$

as $\pi_m = 1 - \pi_1 - \ldots - \pi_{m-1}$.

$$\frac{1}{\pi_1} \sum_{i=1}^n w_{i1}^{(0)} = \ldots = \frac{1}{\pi_{m-1}} \sum_{i=1}^n w_{i,m-1}^{(0)} = c$$

$$\hat{\pi}_k = \frac{\sum_{i=1}^n w_{ik}^{(0)}}{c}$$

$$\pi_j^{(1)} = \frac{\sum_{i=1}^n w_{ij}^{(0)}}{n}$$

And in general,

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

3. Covariance

The asymptotic covariance matrix of $\hat{\boldsymbol{\pi}}$ can be approximated by the inverse of the observed Fisher information matrix. Since there are only $m-1$ free parameters, let $\boldsymbol{\pi}' = (\pi_1, \ldots, \pi_{m-1})$. The likelihood can then be expressed as:

$$l(\boldsymbol{\pi}') = \sum_{i=1}^{n} \log \left\{ \left(\sum_{j=1}^{m-1} \pi_j f_j \right) + (1 - \pi_1, \ldots, \pi_{m-1}) f_m \right\} \tag{4}$$

The observed information matrix, $I(\boldsymbol{\pi}'|\mathbf{x}, \mathbf{y})$, is given by the $m-1 \times m-1$ negative hessian of (4):

$$I(\boldsymbol{\pi}|\mathbf{x}, \mathbf{y}) = -\frac{\partial^2 l(\boldsymbol{\pi}')}{\partial \boldsymbol{\pi}' \partial \boldsymbol{\pi}'^T} = - \begin{bmatrix} \frac{\partial^2 l(\boldsymbol{\pi}')}{\partial^2 \pi_1} & \frac{\partial^2 l(\boldsymbol{\pi}')}{\partial \pi_1 \partial \pi_2} & \cdots & \frac{\partial^2 l(\boldsymbol{\pi}')}{\partial \pi_1 \partial \pi_{m-1}} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 l(\boldsymbol{\pi}')}{\partial \pi_{m-1} \partial \pi_1} & \frac{\partial^2 l(\boldsymbol{\pi}')}{\partial \pi_{m-1} \partial \pi_2} & \cdots & \frac{\partial^2 l(\boldsymbol{\pi}')}{\partial^2 \pi_{m-1}} \end{bmatrix}$$

where

$$\frac{\partial l(\boldsymbol{\pi}')}{\partial \pi_k} = \sum_{i=1}^{n} \frac{f_k - f_m}{\sum_{j=1}^{m} \pi_j f_j}$$

$$\frac{\partial^2 l(\boldsymbol{\pi}')}{\partial \pi_k \partial \pi_r} = -\sum_{i=1}^{n} \frac{(f_k - f_m)(f_r - f_m)}{(\sum_{j=1}^{g} \pi_j f_j)^2 f_r}$$

The inverse of $I(\boldsymbol{\pi}'|\mathbf{x}, \mathbf{y})$ provides estimates of the variance, covariance, and correlation of $\hat{\boldsymbol{\pi}}$ as

$$\mathrm{Cov}(\hat{\pi}_p, \hat{\pi}_q) = \begin{cases} \left[I^{-1}(\hat{\boldsymbol{\pi}}) \right]_{pq} & p, q < m \\ -\sum\limits_{j=1}^{m-1} \mathrm{Cov}(\hat{\pi}_j, \hat{\pi}_q) & p = m, q < m \\ \sum\limits_{j=1}^{m-1} \sum\limits_{k=1}^{m-1} \mathrm{Cov}(\hat{\pi}_j, \hat{\pi}_q) & p, q = m \end{cases}$$

$$\mathrm{Var}(\hat{\pi}_j) = \sigma_j^2 = \left\{ \mathrm{Cov}(\hat{\boldsymbol{\pi}}) \right\}_{jj}$$

$$\mathrm{Corr}(\hat{\pi}_p, \hat{\pi}_q) = \frac{\mathrm{Cov}(\hat{\pi}_p, \hat{\pi}_q)}{\sqrt{\sigma_p^2 \sigma_q^2}}$$

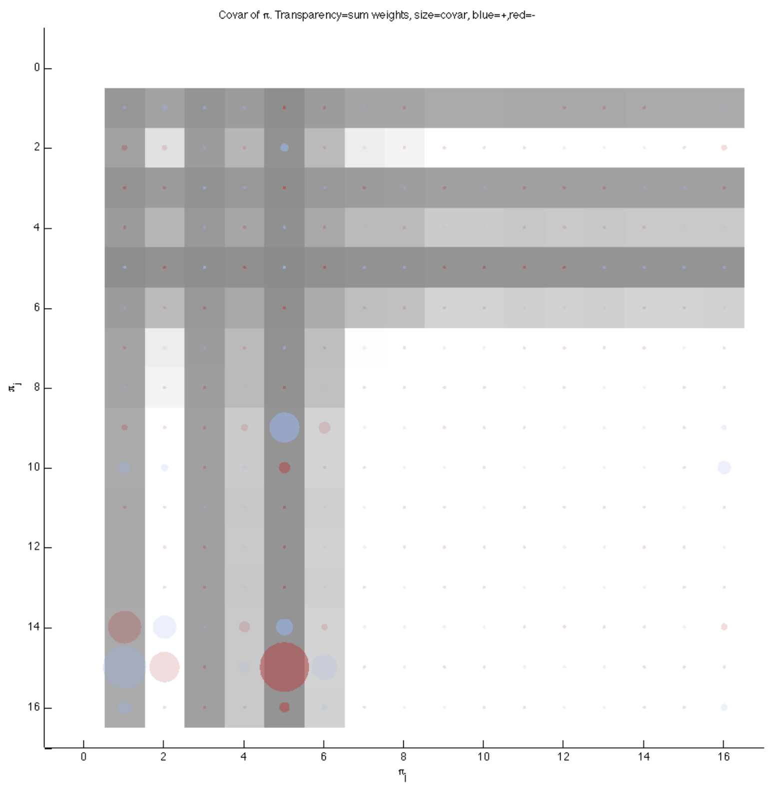
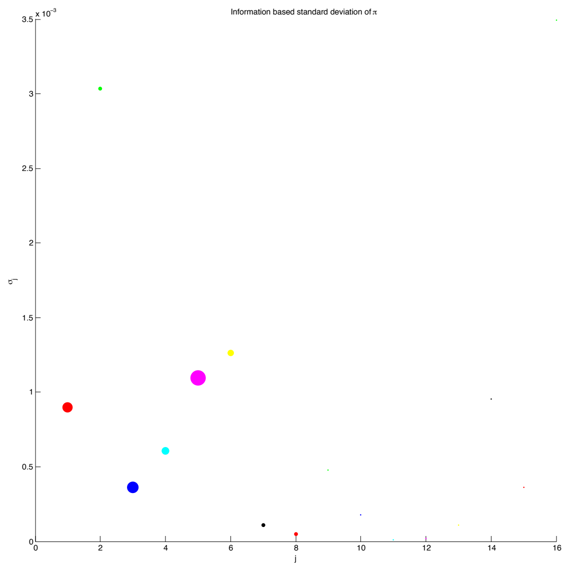


TABLE 1. Covariance, in percent, of $\hat{\boldsymbol{\pi}}$ overlaid on $\hat{\pi}_i + \hat{\pi}_j$. Larger dots are covariances; negative covariances are red.

0.0001	0.001	-	-	-0.0004	-0.0001	-	-	0.0001	-	-	-	-	-	-	0.0007
-0.001	-0.0009	-	-0.0002	0.0015	-0.0004	-	-	-	-	-	-	-	-	-	-0.001
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-0.0002	0.0003	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-	-0.0001	-	-	0.0001	-	-	-	-	-	-	-	-	-	-	-
0.0003	-0.0001	-	0.0001	-0.0004	0.0002	-	-	-	-	-	-	-	-	-	-
-0.0001	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
-0.0011	-0.0001	-	-0.0013	0.0053	-0.0021	-	-	-	-	-	-	-	0.0001	-	0.001
0.0023	0.0013	-0.0001	0.0008	-0.002	-	-0.0001	0.0001	-	-	-	-	-	-	-	0.0023
-	-	-	-	-0.0001	0.0001	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
0.0003	-0.0003	-	-	-0.0001	0.0001	-	-	-	-	-	-	-	-	-	-
-0.0058	0.0041	0.0001	-0.0019	0.003	-0.0011	0.0002	0.0001	0.0003	-	-	-0.0001	-0.0001	-0.0001	-	-0.0012
0.0076	-0.0053	-0.0001	0.0019	-0.0086	0.0045	-0.0003	0.0002	-0.0003	-	-	-	-	-0.0001	-	-0.0005
0.0025	-0.0001	-0.0001	-0.0005	-0.0018	0.0011	-0.0001	0.0003	-	-	-	-	-	-	-	0.0012



(A) Larger dots represent more weight accounted for by that $\hat{\pi}_j$.

	$\hat{\pi}_j$	π_j	$\sigma_{\hat{\pi}_j}$	Z-score
1	16.04	14.467	0.0898	17.509
2	3.320	7.435	0.3033	-13.564
3	21.392	20.991	0.0361	11.067
4	8.846	9.234	0.0608	-6.369
5	36.443	33.656	0.1094	25.479
6	7.882	8.019	0.1264	-1.080
7	2.560	2.400	0.0108	14.793
8	2.239	2.573	0.0047	-70.08
9	0	0.118	0.0480	-2.474
10	0	0.076	0.0178	-4.317
11	0.421	0.358	0.0009	65.245
12	0.246	0.253	0.0021	-2.959
13	0.485	0.225	0.0110	23.56
14	0.008	0.048	0.0954	-0.420
15	0.030	0.024	0.0362	0.183
16	0.083	0.046	0.3494	0.103

(B)

FIGURE 1. Variance of $\hat{\boldsymbol{\pi}}$, true $\boldsymbol{\pi}$, in percent, and z-scores.

$$\text{Z-score} = \frac{\hat{\pi}_j - \pi}{\sigma_{\hat{\pi}_j}}$$

4. Likelihood ratio test

Given certain conditions

$$H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_{\text{true}}$$
$$H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}_{\text{true}}$$

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

$$\begin{aligned} \Lambda_{\text{Halo 3}} &= 25.025 \sim \chi^2_{15} \\ \text{p-value 10k} &= 4.961\% \\ \text{p-value 30k} &= 1.3 \times 10^{-5}\% \\ \text{p-value 50k} &= 1.4 \times 10^{-12}\% \end{aligned}$$

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.