

METALLICITY

NOVEMBER 16, 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

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

where

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

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

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

Thus the likelihood of (1) is

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

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

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

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

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

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

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 $\text{argmax}_{\boldsymbol{\pi}}$ 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. 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 .

$$(3) \quad \mathbb{E}_{\boldsymbol{\pi}} \left[\ell(\boldsymbol{\pi}) | \mathbf{x}, \mathbf{y} \right] = \sum_{i=1}^n \sum_{j=1}^m \mathbb{E}_{\boldsymbol{\pi}} [z_{ij} | x_i, y_i] \{ \log f_j(x_i, y_i) + \log \pi_j \}$$

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

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\pi}} [z_{ij} | x_i, y_i] &= \text{Probability} \left((x_i, y_i) \mapsto G_j | x_i, y_i \right) \\ &= \Pr_{\boldsymbol{\pi}}(z_{ij} | x_i, y_i) \\ &= \frac{p(x_i, y_i | z_{ij} = 1) p(z_{ij} = 1)}{p(x_i, y_i)} \end{aligned}$$

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

$$(4) \quad \mathbb{E}_{\boldsymbol{\pi}}[z_{ij}|x_i, y_i] = \frac{\pi_j f_j(x_i, y_i)}{\sum_{j=1}^m \pi_j f_j(x_i, y_i)}$$

To iteratively evaluate this expectation, we let $w_{ij}^{(t)}$ be (4) at the t^{th} step:

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

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. Maximization step. We now have an explicit formulation for the expected log likelihood (4) given a single parameter $\boldsymbol{\pi}$, plus the data. The argument of the maximum of (4) at each iteration t provides an estimate that approaches the MLE of $\boldsymbol{\pi}$, and is given by:

$$(5) \quad \hat{\boldsymbol{\pi}}^{(t)} = \underset{\boldsymbol{\pi}}{\operatorname{argmax}} \mathbb{E}[\ell(\boldsymbol{\pi})|\mathbf{x}, \mathbf{y}, \hat{\boldsymbol{\pi}}^{(t-1)}]$$

Accounting for the $m-1$ free parameters of $\boldsymbol{\pi}$, differentiation of (4) proceeds, for $k = 1, \dots, m-1$, as:

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

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

Consequently, using some constant, c , we must have

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

The unknown constant c appears problematic, but, because $\sum_{j=1}^m \pi_j = 1$, algebraic manipulation reveals that $c = n$, yielding a final solution that can be numerically evaluated:

$$\hat{\pi}_k^{(t)} = \frac{\sum_{i=1}^n w_{ij}^{(t-1)}}{n}$$

$$\hat{\pi}_m^{(t)} = 1 - \pi_1 - \dots - \pi_{m-1}$$

In our case, computation of $\hat{\pi}$ converges relatively quickly for all starting values: on the order of 600 iterations, or half a minute, for our stopping criteria. Large π_k values typically emerge after two or three iterations, and most change, absolutely speaking, occurs in the first fifty to one hundred iterations.

3. COVARIANCE AND CORRELATION OF $\hat{\boldsymbol{\pi}}$

The asymptotic covariance matrix of $\hat{\boldsymbol{\pi}}$ can be approximated by the inverse of the observed Fisher information matrix, I .

As $\pi_m = 1 - \sum_{j=1}^{m-1} \pi_j$, there are only $m - 1$ free parameters. Thus let $\boldsymbol{\pi}' = (\pi_1, \dots, \pi_{m-1})$. Using $f_{ij} = f_j(x_i, y_i)$ for brevity, the likelihood can then be expressed as:

$$(6) \quad \ell(\boldsymbol{\pi}') = \sum_{i=1}^n \log \left\{ \left(\sum_{j=1}^{m-1} \pi_j f_{ij} \right) + (1 - \pi_1, \dots, \pi_{m-1}) f_{im} \right\}$$

The observed information matrix, I , is the $m - 1 \times m - 1$ negative hessian of (6), evaluated at the observed data points:

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

where

$$\frac{\partial \ell(\boldsymbol{\pi}')}{\partial \pi_k} = \sum_{i=1}^n \frac{f_{ik} - f_{im}}{\sum_{j=1}^m \pi_j f_{ij}} \quad \text{and} \quad \frac{\partial^2 \ell(\boldsymbol{\pi}')}{\partial \pi_k \partial \pi_r} = - \sum_{i=1}^n \frac{(f_{ik} - f_{im})(f_{ir} - f_{im})}{(\sum_{j=1}^g \pi_j f_{ij})^2}$$

The observed information derived covariance matrix of $\boldsymbol{\pi}'$ yields the following estimates for covariance and correlation for all m estimated weights in $\hat{\boldsymbol{\pi}}$:

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

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

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

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

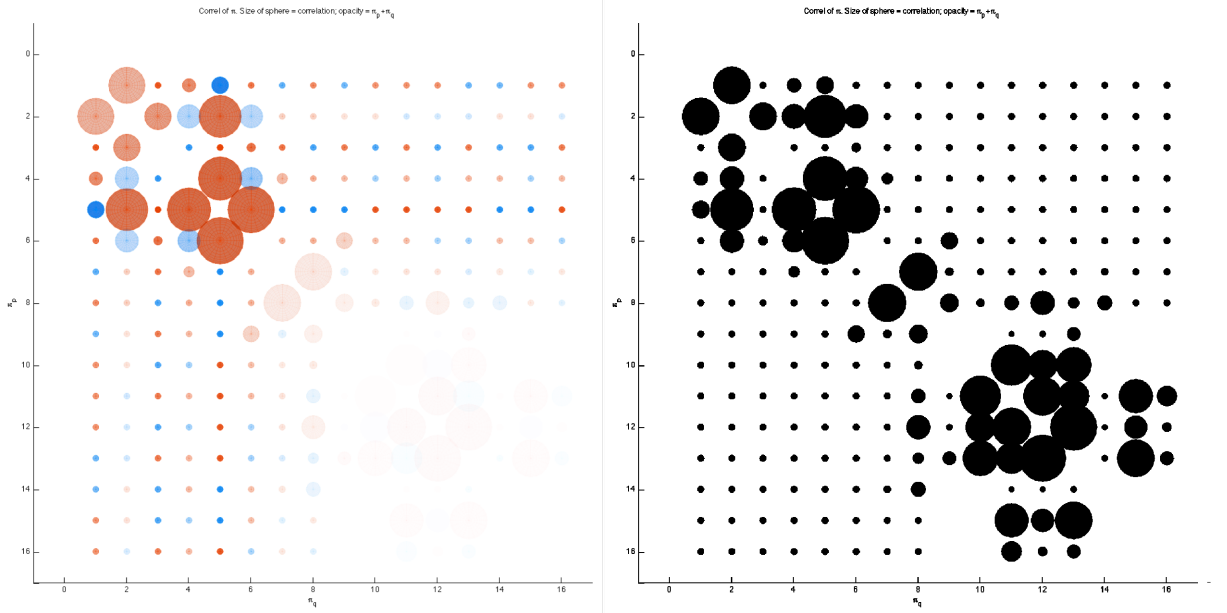


FIGURE 1. Left: correlation where size of sphere represents correlation value, and opacity of sphere represents $\pi_q + \pi_p$. Orange bubbles represent negative correlation, and blue positive. Right: Same graph, but without transparency.

TABLE 1. Correlation matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1	-0.592	-0.026	-0.219	0.274	-0.101	0.005	-0.008	0.007	-0.004	-0.003	-0.006	0.003	0.008	-0.001	-0.002
2	-0.592	1	-0.436	0.385	-0.685	0.378	-0.03	-0.006	-0.026	-0.003	0.002	0.002	0.003	-0.029	-0.004	0.006
3	-0.026	-0.436	1	0.074	-0.008	-0.141	-0.036	0.008	-0.006	0.004	-0.006	0.004	-0.008	0.025	0.002	-0.009
4	-0.219	0.385	0.074	1	-0.706	0.368	-0.173	-0.021	-0.045	0.013	-0.012	0.051	-0.022	-0.01	0.002	-0.003
5	0.274	-0.685	-0.008	-0.706	1	-0.757	0.057	0.008	0.083	-0.002	0	-0.018	-0.006	0.051	0.002	-0.001
6	-0.101	0.378	-0.141	0.368	-0.757	1	-0.012	-0.07	-0.264	-0.007	-0.01	0.027	0.02	-0.028	-0.009	0.01
7	0.005	-0.03	-0.036	-0.173	0.057	-0.012	1	-0.602	0.129	-0.013	-0.065	-0.023	0.025	-0.109	0.021	-0.002
8	-0.008	-0.006	0.008	-0.021	0.008	-0.07	-0.602	1	-0.29	-0.125	0.226	-0.381	0.173	0.232	-0.082	0.049
9	0.007	-0.026	-0.006	-0.045	0.083	-0.264	0.129	-0.29	1	0.112	-0.072	0.111	-0.211	-0.751	0.095	-0.048
10	-0.004	-0.003	0.004	0.013	-0.002	-0.007	-0.013	-0.125	0.112	1	-0.665	0.488	-0.567	0.091	0.439	-0.54
11	-0.003	0.002	-0.006	-0.012	0	-0.01	-0.065	0.226	-0.072	-0.665	1	-0.62	0.502	-0.044	-0.542	0.326
12	-0.006	0.002	0.004	0.051	-0.018	0.027	-0.023	-0.381	0.111	0.488	-0.62	1	-0.748	-0.006	0.376	-0.156
13	0.003	0.003	-0.008	-0.022	-0.006	0.02	0.025	0.173	-0.211	-0.567	0.502	-0.748	1	0.015	-0.599	0.218
14	0.008	-0.029	0.025	-0.01	0.051	-0.028	-0.109	0.232	-0.751	0.091	-0.044	-0.006	0.015	1	0.008	-0.248
15	-0.001	-0.004	0.002	0.002	0.002	-0.009	0.021	-0.082	0.095	0.439	-0.542	0.376	-0.599	0.008	1	-0.382
16	-0.002	0.006	-0.009	-0.003	-0.001	0.01	-0.002	0.049	-0.048	-0.54	0.326	-0.156	0.218	-0.248	-0.382	1

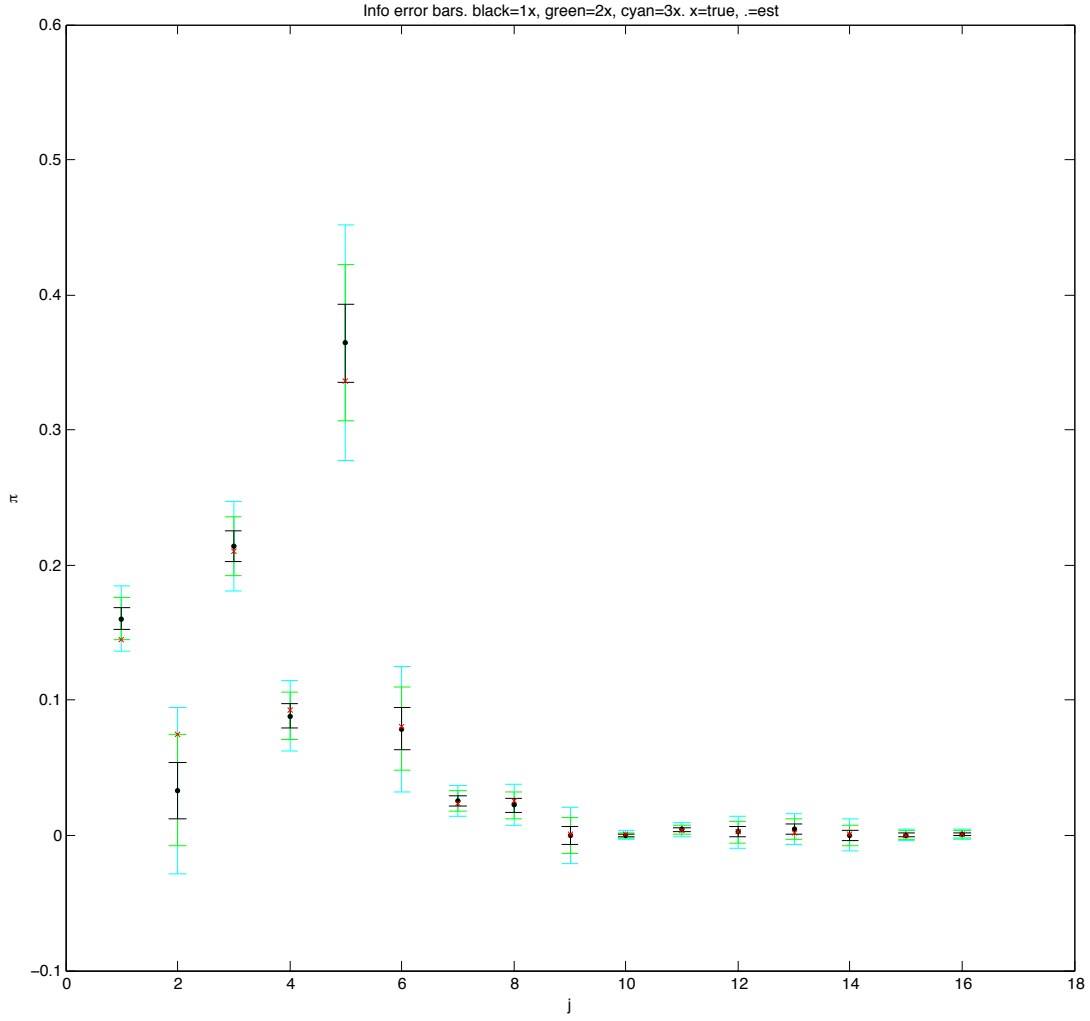


FIGURE 2. π plus information based error bars for $\pm\sigma$ (black), $\pm 2\sigma$ (green), and $\pm 3\sigma$ (cyan). A red \times represents the true value, and a black dot represents the estimated values, $\hat{\pi}$.

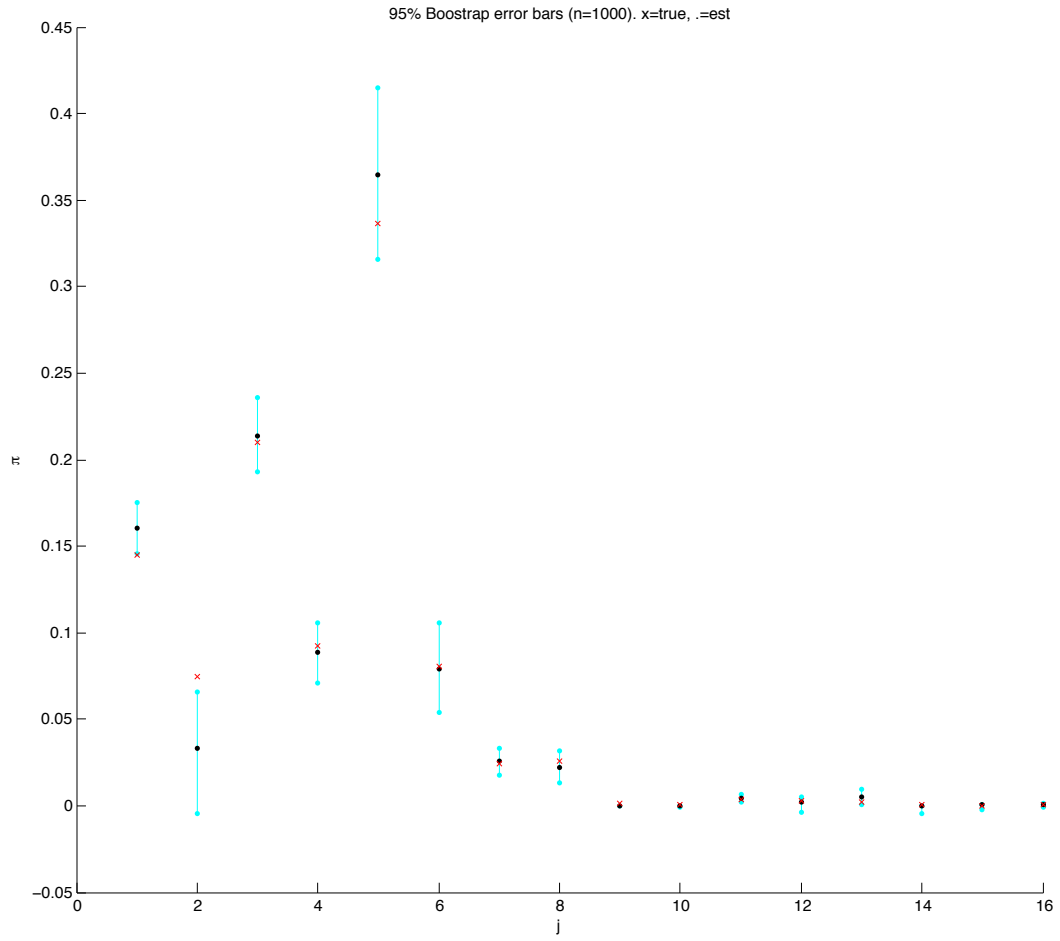


FIGURE 3. M-out-of n bootstrapped confidence intervals for π . A red \times represents the true value, and a black dot represents the estimated values, $\hat{\pi}$.

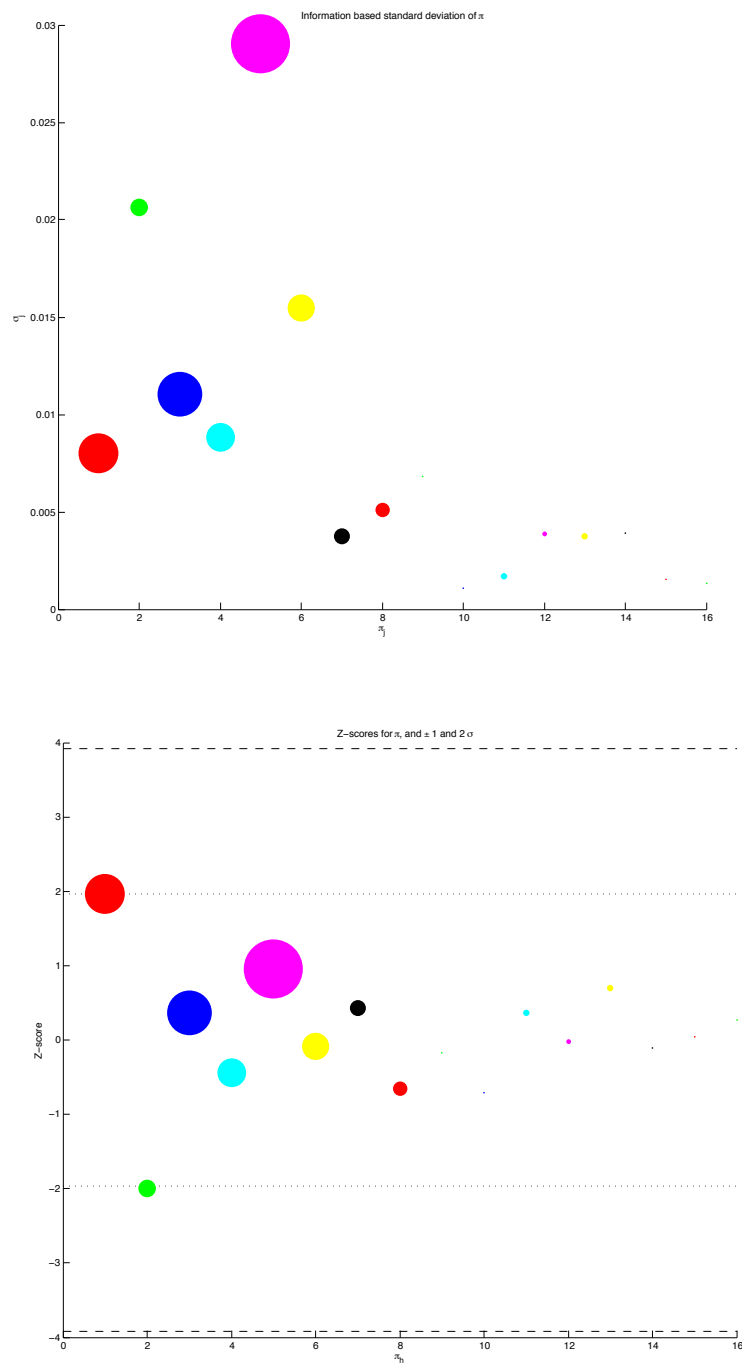


FIGURE 4. Left: standard deviation of each π_j . Right: Z-score, with 1 and 2 standard deviations marked as dotted lines. Colors are same as EM diagnostic plots. Size represents the value of π_j .

	$\hat{\pi}_j$	π_j	Std. dev.	Z-score
1	16.04	14.47	0.0080	1.965
2	3.32	7.44	0.0206	-1.995
3	21.39	20.99	0.0110	0.363
4	8.85	9.23	0.0088	-0.439
5	36.44	33.66	0.0290	0.96
6	7.88	8.02	0.0154	-0.088
7	2.56	2.4	0.0037	0.428
8	2.24	2.57	0.0051	-0.655
9	0	0.12	0.0068	-0.174
10	0	0.08	0.0010	-0.706
11	0.42	0.36	0.0017	0.367
12	0.25	0.25	0.0038	-0.016
13	0.49	0.23	0.0037	0.694
14	0.01	0.05	0.0039	-0.103
15	0.03	0.02	0.0015	0.044
16	0.08	0.05	0.0013	0.27

FIGURE 5. Model 3 EM results

4. LIKELIHOOD RATIO TEST

Given certain regularity conditions, let

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

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

The likelihood ratio test is then

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

For halo 3,

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

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

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

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

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.