Bayesian Computation in Color-Magnitude Diagrams

SA, AA, PT, EE, MCMC and ASIS in CMDs

Paul Baines

Department of Statistics
Harvard University

October 19, 2009

Overview

- Motivation and Introduction
- Modelling Color-Magnitude Diagrams
- Bayesian Statistical Computation
- Acronyms in SC: MH, ASIS, EE, PT

Stellar Populations

We want to study stars, or ‘clusters’ of stars, and are interested in their mass and age, and sometimes their metallicity.

Stellar Populations

We want to study stars, or 'clusters' of stars, and are interested in their mass and age, and sometimes their metallicity.

The way this is usually done is by:

I Taking photometric observations of the stars in a number of different wavelengths,
II Compare observed values with what the theory predicts, and,
III Select mass, age and metallicity to minimize the χ^2 statistic.

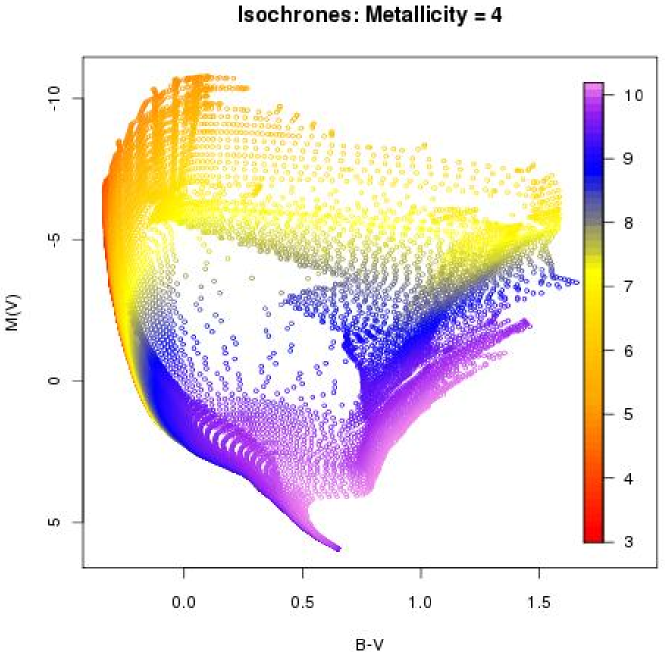
Stellar Populations

We want to study stars, or 'clusters' of stars, and are interested in their mass and age, and sometimes their metallicity.

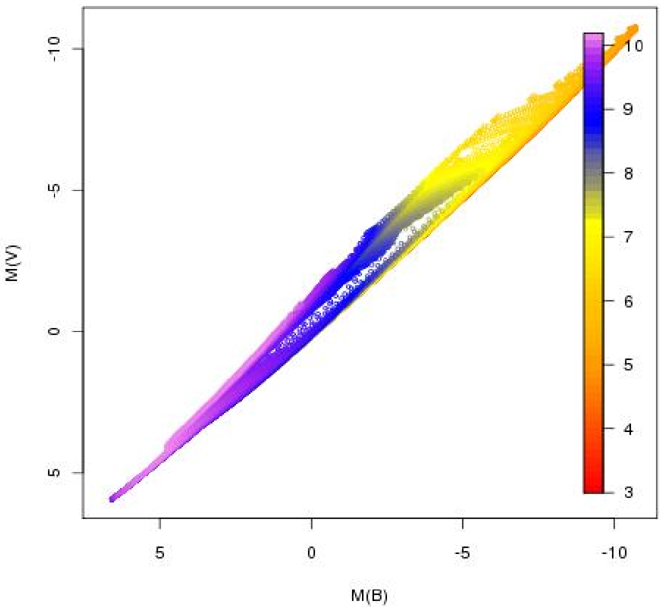
The way this is usually done is by:

I Taking photometric observations of the stars in a number of different wavelengths,
II Compare observed values with what the theory predicts, and,
III Select mass, age and metallicity to minimize the χ^2 statistic.

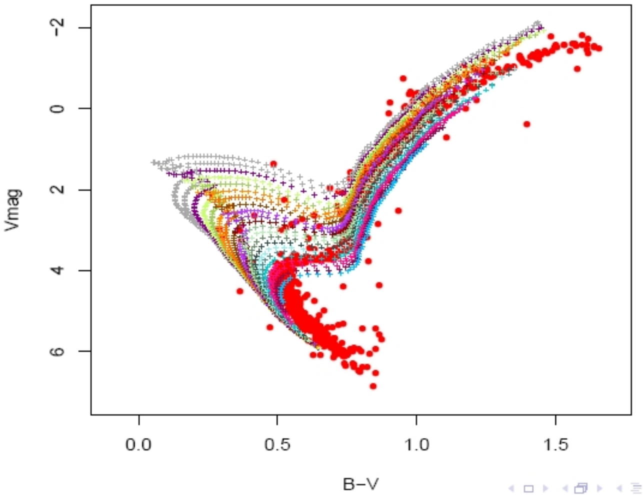
. . . doesn't utilize any knowledge of underlying structure, propagation of error, hard to extend coherently. . .


Isochrones: Metallicity = 4
M(V)
-10
-5
0
5
0.0
0.5
1.0
1.5
B-V
10
9
8
7
6
5
4
3

Isochrones: Metallicity = 4



NGC 104 (47 Tuc)



Gaussian Errors

The observed magnitudes in different bands would ideally match the theoretical isochrone values, but there is both (i) measurement error, and, (ii) natural variability (photon arrivals).

The uncertainty at this stage is treated as Gaussian, although these can be correlated across bands.

Gaussian Errors

The observed magnitudes in different bands would ideally match the theoretical isochrone values, but there is both (i) measurement error, and, (ii) natural variability (photon arrivals).

The uncertainty at this stage is treated as Gaussian, although these can be correlated across bands.

(Mechanically unimportant)

The Likelihood

Assuming gaussian errors with common correlation matrix:

$$y_i \Big| A_i, M_i, Z \sim N\left(\tilde{f}_i, \mathbf{R}\right) \qquad i = 1, \ldots, n \tag{1}$$

Where,

$$y_i = \begin{pmatrix} \frac{1}{\sigma_i^{(B)}} B_i \\ \frac{1}{\sigma_i^{(V)}} V_i \\ \frac{1}{\sigma_i^{(I)}} I_i \end{pmatrix}, \tilde{f}_i = \begin{pmatrix} \frac{1}{\sigma_{Bi}} \cdot f_b(M_i, A_i, Z) \\ \frac{1}{\sigma_{Vi}} \cdot f_v(M_i, A_i, Z) \\ \frac{1}{\sigma_{Ii}} \cdot f_i(M_i, A_i, Z) \end{pmatrix}, \mathbf{R} = \begin{pmatrix} 1 & \rho^{(BV)} & \rho^{(BI)} \\ \rho^{(BV)} & 1 & \rho^{(VI)} \\ \rho^{(BI)} & \rho^{(VI)} & 1 \end{pmatrix} .$$

The Likelihood

Assuming gaussian errors with common correlation matrix:

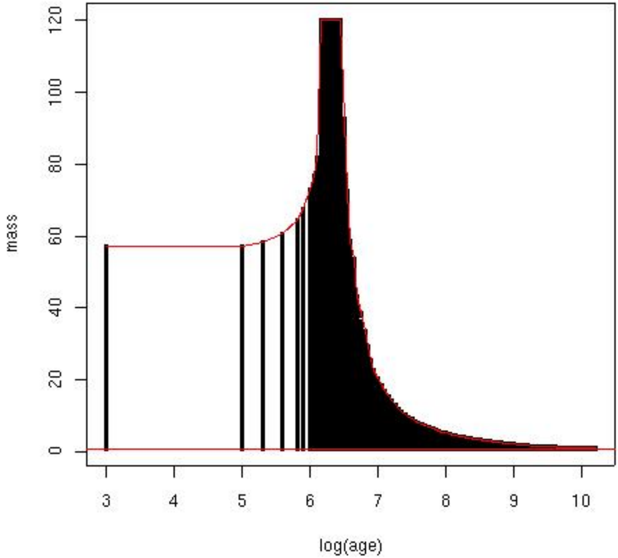
$$y_i \middle| A_i, M_i, Z \sim N\left(\tilde{f}_i, \mathbf{R}\right) \qquad i = 1, \ldots, n \tag{1}$$

Where,

$$y_i = \begin{pmatrix} \frac{1}{\sigma_i^{(B)}} B_i \\ \frac{1}{\sigma_i^{(V)}} V_i \\ \frac{1}{\sigma_i^{(I)}} I_i \end{pmatrix}, \tilde{f}_i = \begin{pmatrix} \frac{1}{\sigma_{Bi}} \cdot f_b(M_i, A_i, Z) \\ \frac{1}{\sigma_{Vi}} \cdot f_v(M_i, A_i, Z) \\ \frac{1}{\sigma_{Ii}} \cdot f_i(M_i, A_i, Z) \end{pmatrix}, \mathbf{R} = \begin{pmatrix} 1 & \rho^{(BV)} & \rho^{(BI)} \\ \rho^{(BV)} & 1 & \rho^{(VI)} \\ \rho^{(BI)} & \rho^{(VI)} & 1 \end{pmatrix}.$$

Now, we assume a hierarchical Bayesian model on the mass, age, and metallicities.

Isochrone table layout
(support)



Joint Distribution of Mass and Age

Even before we observe any data, we know that the distributions of stellar mass and age are not independent. We know *a priori* that, (for a star to be in the population of all possibly observable stars), old stars cannot have large mass, likewise for very young stars. Hence, we specify the prior on mass conditional on age:

$$p\left(M_i | A_i, M_{min}, M_{max}\left(A_i\right), \alpha\right) \propto \frac{1}{M_i^{\alpha}} \cdot \mathbf{1}_{\{M_i \in [M_{min}, M_{max}(A_i)]\}} \qquad (2)$$

i.e. $M_i | A_i, M_{min}, M_{max}\left(A_i\right), \alpha \sim \text{Truncated-Pareto}$.

Age

For age we assume the following hierarchical structure:

$$A_i|\mu_A, \sigma_A^2 \stackrel{iid}{\sim} N\left(\mu_A, \sigma_A^2\right) \tag{3}$$

where $A_i = \log_{10}(\mathrm{Age})$, with μ_A and σ_A^2 hyperparameters...

Age

For age we assume the following hierarchical structure:

$$A_i|\mu_A, \sigma_A^2 \stackrel{iid}{\sim} N\left(\mu_A, \sigma_A^2\right) \tag{3}$$

where $A_i = \log_{10}(\mathrm{Age})$, with μ_A and σ_A^2 hyperparameters...

Idea: Allow greater flexibility/reduce dimensionality, investigate strength of assumptions, capture underlying structure, correlation etc

Hyperparameters

Next, we model the hyperparameters with the simple conjugate form:

$$\mu_A|\sigma_A^2 \sim N\left(\mu_0, \frac{\sigma_A^2}{\kappa_0}\right), \qquad \sigma_A^2 \sim Inv-\chi^2\left(\nu_0, \sigma_0^2\right) \tag{4}$$

Where μ_0, κ_0, ν_0 and σ_0^2 are fixed by the user to represent prior knowledge (or lack of).

Sensitivity (or robustness) to the prior is of great interest.

Correlation

We assume a uniform prior over the space of positive definite correlation matrices.

This isn't quite uniform on each of $\rho^{(BV)}, \rho^{(BI)}$ and $\rho^{(VI)}$, but it is close for low dimensions.

Putting it all together

$$y_i = \begin{pmatrix} \frac{1}{\sigma_i^{(B)}} B_i \\ \frac{1}{\sigma_i^{(V)}} V_i \\ \frac{1}{\sigma_i^{(I)}} I_i \end{pmatrix} \Bigg| A_i, M_i, Z \sim N\left(\tilde{f}_i, \mathbf{R}\right) \qquad i = 1, \ldots, n.$$

$$M_i | A_i, M_{min}, \alpha \sim \text{Truncated-Pareto}\left(\alpha - 1, M_{min}, M_{max}\left(A_i\right)\right)$$

$$A_i | \mu_A, \sigma_A^2 \overset{iid}{\sim} N\left(\mu_A, \sigma_A^2\right)$$

$$\mu_A | \sigma_A^2 \sim N\left(\mu_0, \frac{\sigma_A^2}{\kappa_0}\right), \qquad \sigma_A^2 \sim Inv - \chi^2\left(\nu_0, \sigma_0^2\right)$$

$$p(\mathbf{R}) \propto \mathbf{1}_{\{\mathbf{R} p.d.\}}$$

MCMC Toolbox

(Almost) everything must use a Gibbs sampler (or conditional maximization) if it has lots of parameters. Gibbs is great but its failings are well documented. Three top-of-the-range tools for building your MCMC scheme:

MCMC Toolbox

(Almost) everything must use a Gibbs sampler (or conditional maximization) if it has lots of parameters. Gibbs is great but its failings are well documented. Three top-of-the-range tools for building your MCMC scheme:

1. Parallel Tempering (Geyer, 1991)

MCMC Toolbox

(Almost) everything must use a Gibbs sampler (or conditional maximization) if it has lots of parameters. Gibbs is great but its failings are well documented. Three top-of-the-range tools for building your MCMC scheme:

1. Parallel Tempering (Geyer, 1991)
2. Equi-Energy Sampling (Kou, Zhou & Wong, 2006)

MCMC Toolbox

(Almost) everything must use a Gibbs sampler (or conditional maximization) if it has lots of parameters. Gibbs is great but its failings are well documented. Three top-of-the-range tools for building your MCMC scheme:

1. Parallel Tempering (Geyer, 1991)
2. Equi-Energy Sampling (Kou, Zhou & Wong, 2006)
3. Ancillarity-Sufficiency Interweaving (ASIS) (Yu & Meng, 2009)

MCMC Toolbox

(Almost) everything must use a Gibbs sampler (or conditional maximization) if it has lots of parameters. Gibbs is great but its failings are well documented. Three top-of-the-range tools for building your MCMC scheme:

1. Parallel Tempering (Geyer, 1991)
2. Equi-Energy Sampling (Kou, Zhou & Wong, 2006)
3. Ancillarity-Sufficiency Interweaving (ASIS) (Yu & Meng, 2009)

We now take a 'hands on' look at these in developing a robust algorithm for analysis Color-Magnitude data.

Default MCMC

Algorithm

Let $\Theta^{(t)} = \left\{ (M_i^{(t)}, A_i^{(t)}) : i = 1, \ldots, n \right\}$.

0. *Set* $t = 0$*. Choose starting states* $\Theta^{(0)}$, $\mu_a^{(0)}$, $\sigma_a^{(0)}$.
1. *Draw* $\Theta^{(t+1)}$ *from* $\{M_i, A_i\}_{i=1,\ldots,n} | \mu_a^{(t)}, \sigma_a^{(t)}, \mathbf{Y}$.
2. *Draw* $\mu_a^{(t+1)}, \sigma_a^{(t+1)}$ *from* $\mu_a, \sigma_a | \Theta^{(t+1)}, \mathbf{Y}$.
3. *Increment* $t \mapsto t + 1$*, return to 1.*

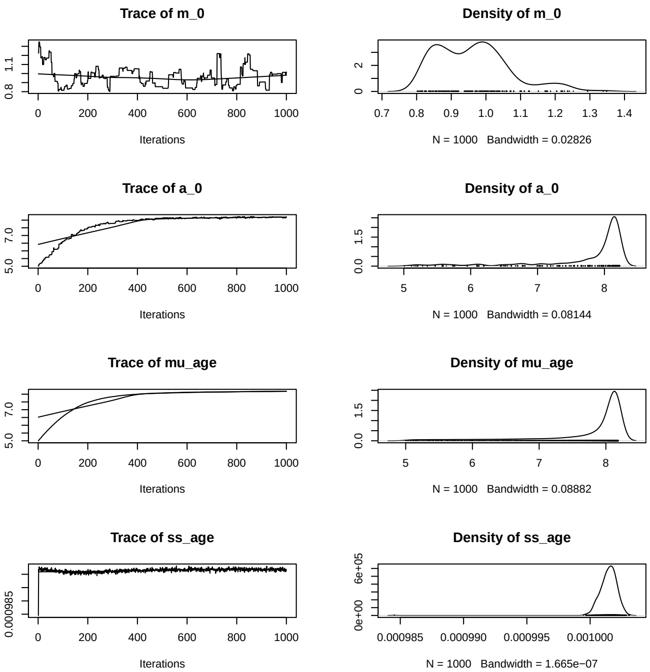
Default MCMC

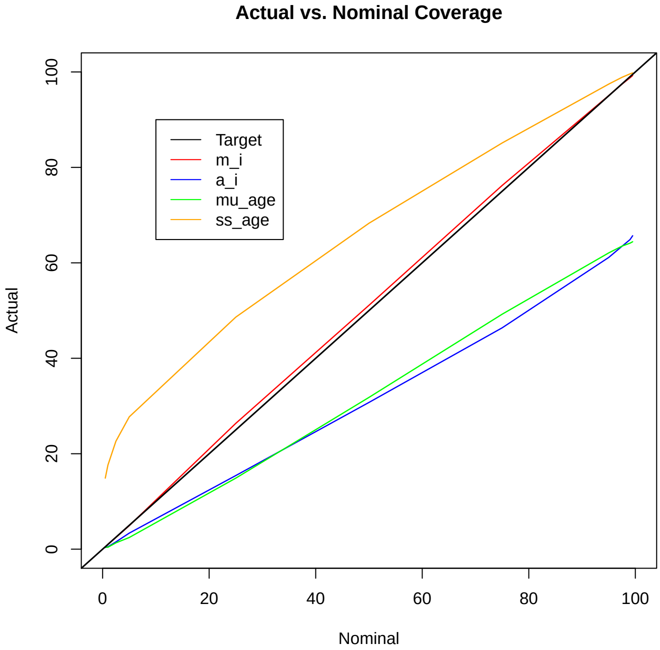
Algorithm

Let $\Theta^{(t)} = \left\{ (M_i^{(t)}, A_i^{(t)}) : i = 1, \ldots, n \right\}$.

0. *Set* $t = 0$. *Choose starting states* $\Theta^{(0)}$, $\mu_a^{(0)}$, $\sigma_a^{(0)}$.
1. *Draw* $\Theta^{(t+1)}$ *from* $\{M_i, A_i\}_{i=1,\ldots,n} | \mu_a^{(t)}, \sigma_a^{(t)}, \mathbf{Y}$.
2. *Draw* $\mu_a^{(t+1)}, \sigma_a^{(t+1)}$ *from* $\mu_a, \sigma_a | \Theta^{(t+1)}, \mathbf{Y}$.
3. *Increment* $t \mapsto t + 1$, *return to 1.*

The samples $(\Theta^{(t)}, \mu_a^{(t)}, \sigma_a^{(t)})$, for $t = B + 1, \ldots, B + N$ form a sample from the target distribution (approximately)...

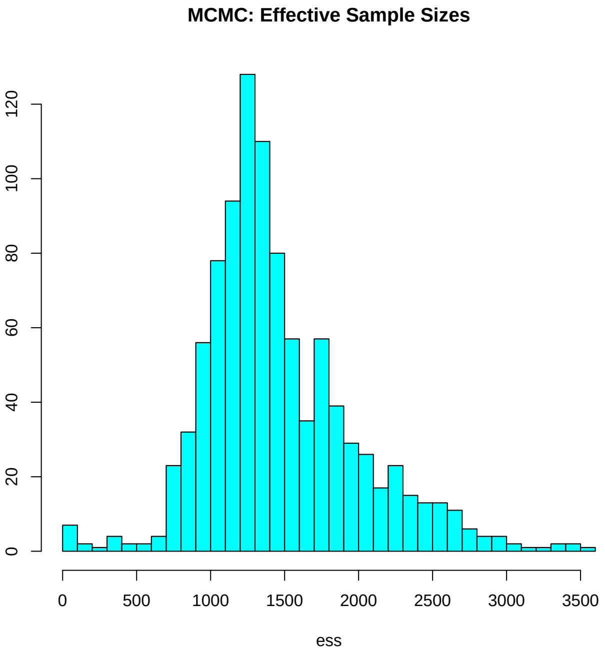

Trace of m_0
Iterations
Density of m_0
N = 1000 Bandwidth = 0.02826
Trace of a_0
Iterations
Density of a_0
N = 1000 Bandwidth = 0.08144
Trace of mu_age
Iterations
Density of mu_age
N = 1000 Bandwidth = 0.08882
Trace of ss_age
Iterations
Density of ss_age
N = 1000 Bandwidth = 1.665e−07


Actual vs. Nominal Coverage
Target
m_i
a_i
mu_age
ss_age
Actual
Nominal
0
20
40
60
80
100
0
20
40
60
80
100


mu_age: Posterior medians vs. Truth
Median(mu_age)
8.6
8.8
9.0
9.2
9.10
9.15
9.20
9.25
9.30
True(mu_age)

	1%	2.5%	5%	25%	50%	75%	95%	97.5%	99%
m_i	1.0	2.5	4.9	26.4	51.1	76.2	95.0	97.5	98.8
a_i	0.5	1.5	3.4	15.4	30.8	46.4	61.1	63.5	64.9
μ_a	0.4	1.3	2.4	14.9	31.8	49.2	62.1	63.5	64.1
σ_a^2	17.6	22.6	27.7	48.6	68.3	85.1	97.5	98.9	99.6

TABLE: Coverage table for the standard SA sampler and synthetic isochrones


MCMC: Effective Sample Sizes
0
20
40
60
80
100
120
0
500
1000
1500
2000
2500
3000
3500
ess

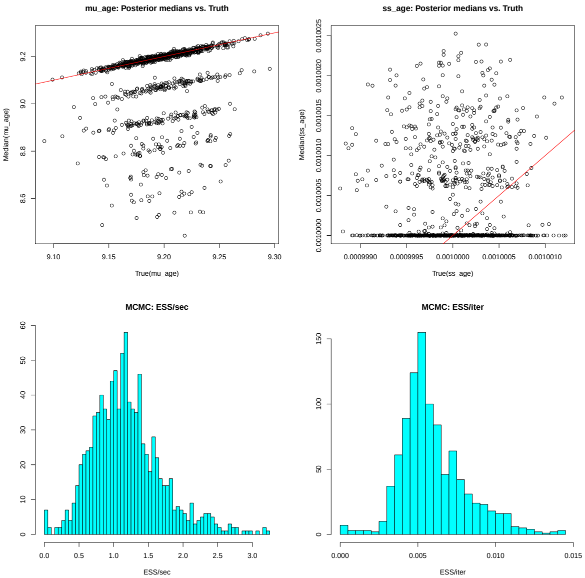


FIGURE: Convergence plots for standard sampler (synthetic isochrones)

Intro to Interweaving

From Yu & Meng (2009): Consider,

$$Y_{obs}|\theta, Y_{mis} \sim N(Y_{mis}, 1) \tag{5}$$

$$Y_{mis}|\theta \sim N(\theta, V). \tag{6}$$

Intro to Interweaving

From Yu & Meng (2009): Consider,

$$Y_{obs}|\theta, Y_{mis} \sim N(Y_{mis}, 1) \quad (5)$$

$$Y_{mis}|\theta \sim N(\theta, V). \quad (6)$$

The standard Gibbs sampler iterates between:

$$Y_{mis}|\theta, Y_{obs} \sim N\left(\frac{\theta + VY_{obs}}{1+V}, \frac{V}{1+V}\right) \quad (7)$$

$$\theta|Y_{mis}, Y_{obs} \sim N\left(Y_{mis}, V\right). \quad (8)$$

Intro to Interweaving

From Yu & Meng (2009): Consider,

$$Y_{obs}|\theta, Y_{mis} \sim N(Y_{mis}, 1) \quad (5)$$

$$Y_{mis}|\theta \sim N(\theta, V). \quad (6)$$

The standard Gibbs sampler iterates between:

$$Y_{mis}|\theta, Y_{obs} \sim N\left(\frac{\theta + VY_{obs}}{1+V}, \frac{V}{1+V}\right) \quad (7)$$

$$\theta|Y_{mis}, Y_{obs} \sim N\left(Y_{mis}, V\right). \quad (8)$$

It is easy to show that the geometric convergence rate for this scheme is $1/(1+V)$ i.e., very slow for small V.

Problems with the SA

By analogy, the very small variance among the stellar ages causes high autocorrelations across the draws and slow convergence.

This standard scheme is known as the **sufficent augmentation (SA)**, since Y_{mis} is a sufficient statistic for μ in the complete-data model.

Ancillary Augmentation

However, by letting $\tilde{Y}_{mis} = Y_{mis} - \theta$ the complete-data model becomes:

$$
\begin{aligned}
Y_{obs} | \tilde{Y}_{mis}, \theta &\sim N(\tilde{Y}_{mis} + \theta, 1), & (9) \\
\tilde{Y}_{mis} &\sim N(0, V). & (10)
\end{aligned}
$$

Ancillary Augmentation

However, by letting $\tilde{Y}_{mis} = Y_{mis} - \theta$ the complete-data model becomes:

$$Y_{obs}|\tilde{Y}_{mis}, \theta \sim N(\tilde{Y}_{mis} + \theta, 1), \quad (9)$$

$$\tilde{Y}_{mis} \sim N(0, V). \quad (10)$$

The Gibbs sampler for this **Ancillary Augmentation (AA)** is:

$$\tilde{Y}_{mis}|\theta, Y_{obs} \sim N\left(\frac{V(Y_{obs} - \theta)}{1 + V}, \frac{V}{1 + V}\right) \quad (11)$$

$$\theta|\tilde{Y}_{mis}, Y_{obs} \sim N\left(Y_{obs} - \tilde{Y}_{mis}, 1\right), \quad (12)$$

with convergence rate $1/(1 + V)$ i.e., extremely fast for small V.

Interweaving

In fact, Yu and Meng (2009) prove that by interweaving the two augmentations, it is possible to produce independent draws in the toy example, and speed up convergence in more general settings.

Interweaving

In fact, Yu and Meng (2009) prove that by interweaving the two augmentations, it is possible to produce independent draws in the toy example, and speed up convergence in more general settings.

The interweaving is done as follows:

- Draw Y_{mis} from the SA: $Y_{mis}^{(t+0.5)}|\theta^{(t)}, Y_{obs}$.

Interweaving

In fact, Yu and Meng (2009) prove that by interweaving the two augmentations, it is possible to produce independent draws in the toy example, and speed up convergence in more general settings.

The interweaving is done as follows:

- Draw Y_{mis} from the SA: $Y_{mis}^{(t+0.5)} | \theta^{(t)}, Y_{obs}$.
- Compute $\tilde{Y}_{mis}^{(t+0.5)} = H(Y_{mis}^{(t+0.5)}; \theta^{(t)})$.

Interweaving

In fact, Yu and Meng (2009) prove that by interweaving the two augmentations, it is possible to produce independent draws in the toy example, and speed up convergence in more general settings.

The interweaving is done as follows:

- Draw Y_{mis} from the SA: $Y_{mis}^{(t+0.5)}|\theta^{(t)}, Y_{obs}$.
- Compute $\tilde{Y}_{mis}^{(t+0.5)} = H(Y_{mis}^{(t+0.5)}; \theta^{(t)})$.
- Draw θ from the AA: $\theta^{(t+0.5)}|\tilde{Y}_{mis}^{(t+0.5)}, Y_{obs}$.

Interweaving

In fact, Yu and Meng (2009) prove that by interweaving the two augmentations, it is possible to produce independent draws in the toy example, and speed up convergence in more general settings.

The interweaving is done as follows:

- Draw Y_{mis} from the SA: $Y_{mis}^{(t+0.5)} | \theta^{(t)}, Y_{obs}$.
- Compute $\tilde{Y}_{mis}^{(t+0.5)} = H(Y_{mis}^{(t+0.5)}; \theta^{(t)})$.
- Draw θ from the AA: $\theta^{(t+0.5)} | \tilde{Y}_{mis}^{(t+0.5)}, Y_{obs}$.
- Compute $Y_{mis}^{(t+1)} = H^{-1}(\tilde{Y}_{mis}^{(t+0.5)}; \theta^{(t+0.5)})$.

Interweaving

In fact, Yu and Meng (2009) prove that by interweaving the two augmentations, it is possible to produce independent draws in the toy example, and speed up convergence in more general settings.

The interweaving is done as follows:

- Draw Y_{mis} from the SA: $Y_{mis}^{(t+0.5)} | \theta^{(t)}, Y_{obs}$.
- Compute $\tilde{Y}_{mis}^{(t+0.5)} = H(Y_{mis}^{(t+0.5)}; \theta^{(t)})$.
- Draw θ from the AA: $\theta^{(t+0.5)} | \tilde{Y}_{mis}^{(t+0.5)}, Y_{obs}$.
- Compute $Y_{mis}^{(t+1)} = H^{-1}(\tilde{Y}_{mis}^{(t+0.5)}; \theta^{(t+0.5)})$.
- Draw θ from the SA: $\theta^{(t+1)} | Y_{mis}^{(t+1)}, Y_{obs}$.

ASIS Sampler

The Gibbs sampler forms a conditional SA, and it can be shown that:

$$\tilde{A}_i = \Phi\left(\frac{A_i - \mu_A}{\sigma_A}\right), \quad \tilde{M}_i = \frac{M_{min}^{-(\alpha-1)} - M_i^{-(\alpha-1)}}{M_{min}^{-(\alpha-1)} - M_{max}(A_i)^{-(\alpha-1)}}, \quad (13)$$

forms an AA:

ASIS Sampler

The Gibbs sampler forms a conditional SA, and it can be shown that:

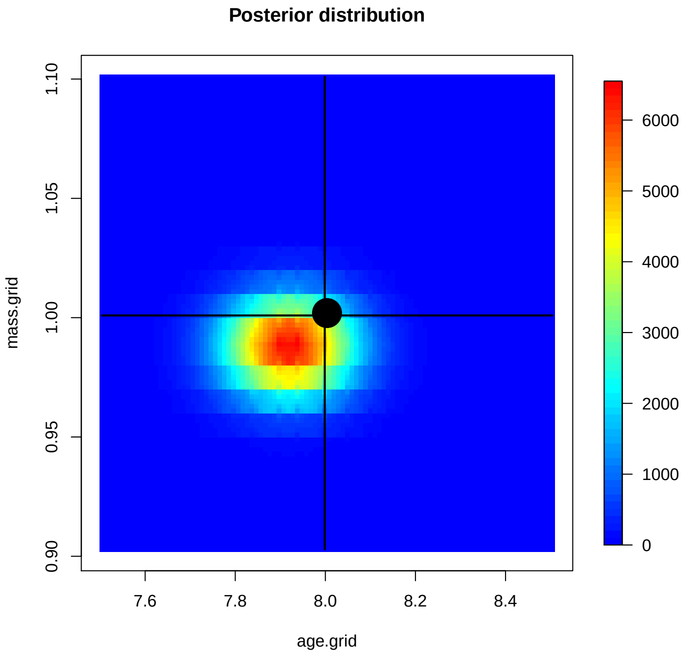
$$\tilde{A}_i = \Phi\left(\frac{A_i - \mu_A}{\sigma_A}\right), \quad \tilde{M}_i = \frac{M_{min}^{-(\alpha-1)} - M_i^{-(\alpha-1)}}{M_{min}^{-(\alpha-1)} - M_{max}(A_i)^{-(\alpha-1)}}, \quad (13)$$

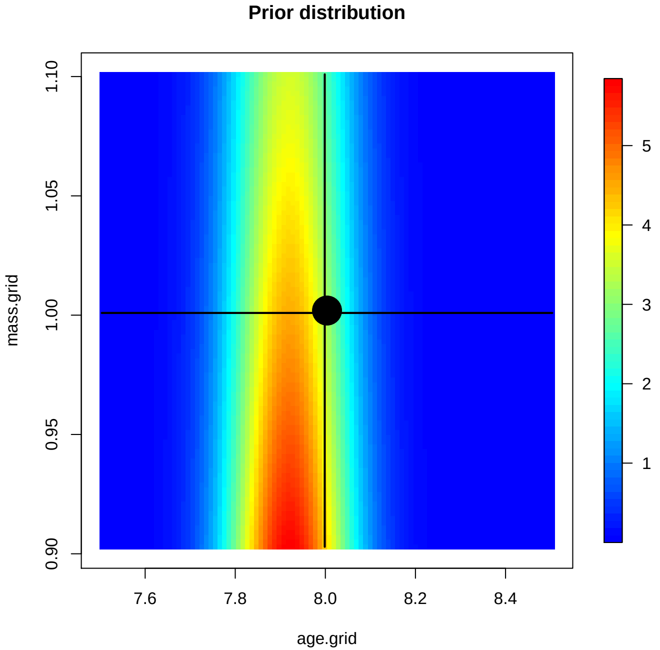
forms an AA:

$$Y_i|\tilde{\mathbf{M}}, \tilde{\mathbf{A}}, \mathbf{R}, \mu_A, \sigma_A^2 \sim N\left(f(M_i(\tilde{M}_i, \tilde{A}_i; \mu_A, \sigma_A), A_i(\tilde{M}_i, \tilde{A}_i; \mu_A, \sigma_A), Z), \mathbf{R}\right)$$

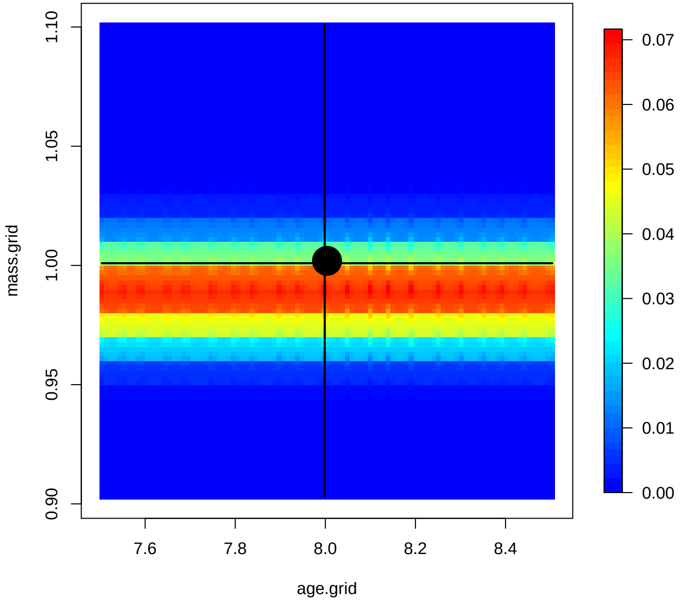
$$\tilde{A}_i|\mu_A, \sigma_A^2 \sim N[0,1], \qquad \tilde{M}_i|\tilde{A}_i, \mu_A, \sigma_A^2 \sim U[0,1]$$

$$\mu_A|\sigma_A^2 \sim N\left(\mu_0, \frac{\sigma_A^2}{\kappa_0}\right), \qquad \sigma_A^2 \sim \text{Inv-}\chi^2\left(\nu_0, \sigma_0^2\right).$$

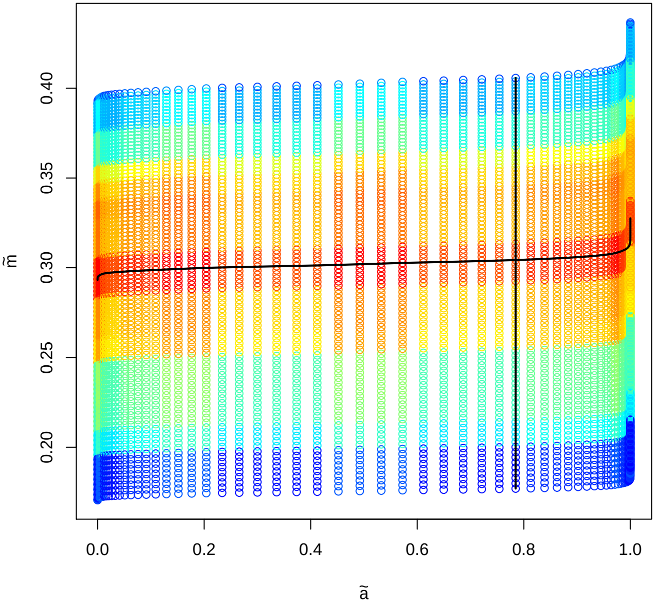

Posterior distribution
mass.grid
1.10
1.05
1.00
0.95
0.90
age.grid
7.6
7.8
8.0
8.2
8.4
6000
5000
4000
3000
2000
1000
0

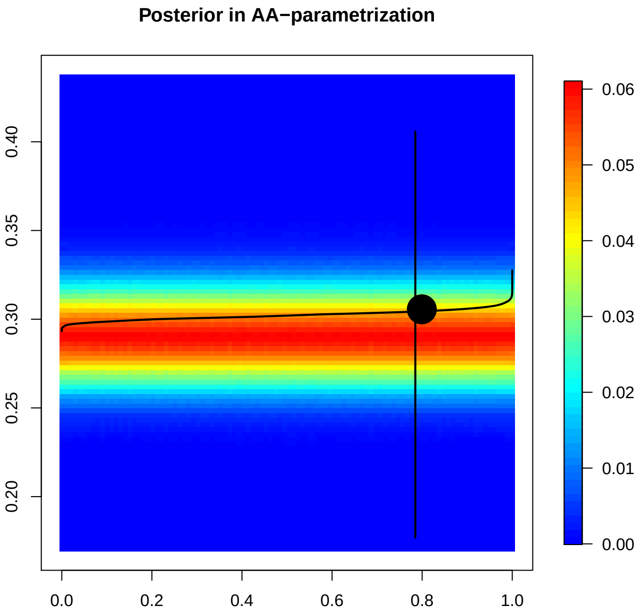

Prior distribution
mass.grid
1.10
1.05
1.00
0.95
0.90
age.grid
7.6
7.8
8.0
8.2
8.4
5
4
3
2
1

Likelihood Surface



Transformation: SA−>AA (transition)





Posterior in AA−parametrization
0.40
0.35
0.30
0.25
0.20
0.0
0.2
0.4
0.6
0.8
1.0
0.06
0.05
0.04
0.03
0.02
0.01
0.00

An ASIS for the CMD Problem

1. Draw $(\mathbf{M}^{(t+0.5)}, \mathbf{A}^{(t+0.5)})$ from SA: $\mathbf{M}, \mathbf{A} | \mathbf{Y}, \mu_A^{(t)}, \sigma_A^{(t)}, \delta^{(t)}$.
2. Compute:
$(\tilde{\mathbf{M}}^{(t+0.5)}, \tilde{\mathbf{A}}^{(t+0.5)}) = H(\mathbf{M}^{(t+0.5)}, \mathbf{A}^{(t+0.5)}; \mu_A^{(t)}, \sigma_A^{(t)})$.
3. Draw $(\mu_A^{(t+0.5)}, \sigma_A^{(t+0.5)})$ from AA:
$\mu_A, \sigma_A^2 | \mathbf{Y}, \tilde{\mathbf{M}}^{(t+0.5)}, \tilde{\mathbf{A}}^{(t+0.5)}, \delta^{(t)}$.
4. Compute:
$(\mathbf{M}^{(t+1)}, \mathbf{A}^{(t+1)}) = H(\tilde{\mathbf{M}}^{(t+0.5)}, \tilde{\mathbf{A}}^{(t+0.5)}; \mu_A^{(t+0.5)}, \sigma_A^{(t+0.5)})$.
5. Redraw $(\mu_A^{(t+1)}, \sigma_A^{(t+1)})$ from SA:
$\mu_A, \sigma_A^2 | \mathbf{Y}, \mathbf{M}^{(t+1)}, \mathbf{A}^{(t+1)}, \delta^{(t)}$.

Beauty and The Beast

The conditional posterior for (mu_a, σ_a^2) under the AA scheme is given by:

$$p(\mu_A, \sigma_A^2 | \tilde{\mathbf{M}}, \tilde{\mathbf{A}}, \mathbf{R}, \mathbf{Y}, \phi) \propto \det(\mathbf{R})^{-n/2} \exp\left\{ -\frac{1}{2} \left[\frac{\kappa_0}{\sigma_A^2} (\mu_A - \mu_0)^2 + \mathrm{tr}\left(\mathbf{R}^{-1} \mathbf{F} \right) \right] \right\}] p(\mu_a, \sigma_a^2)$$

Beauty and The Beast

The conditional posterior for (mu_a, σ_a^2) under the AA scheme is given by:

$$p(\mu_A, \sigma_A^2|\tilde{\mathbf{M}}, \tilde{\mathbf{A}}, \mathbf{R}, \mathbf{Y}, \phi) \propto \det(\mathbf{R})^{-n/2} \exp\left\{-\frac{1}{2}\left[\frac{\kappa_0}{\sigma_A^2}(\mu_A - \mu_0)^2 + \mathrm{tr}\left(\mathbf{R}^{-1}\mathbf{F}\right)\right]\right\}]p(\mu_a, \sigma_a^2)$$

Shifting (μ_a, σ_a^2) and mapping back to the AA will shift every single mass and age, but at a price...

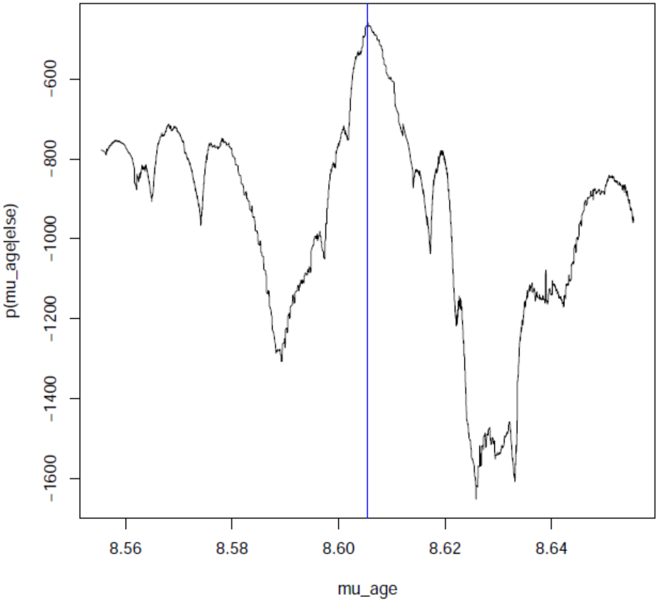


FIGURE: Full conditional log-posterior for μ_a in the AA-sampler.

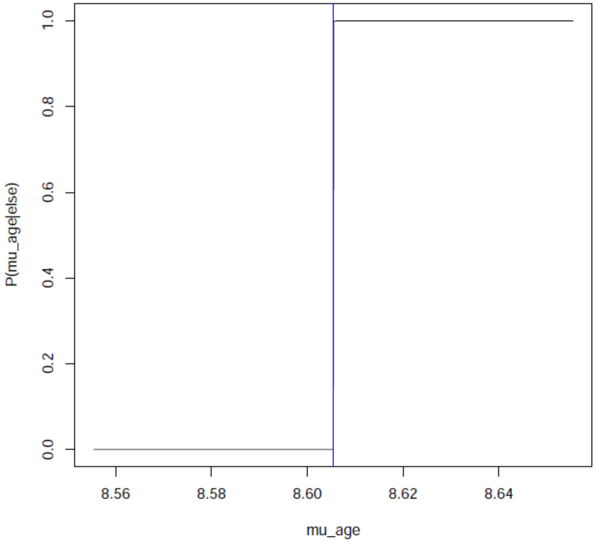


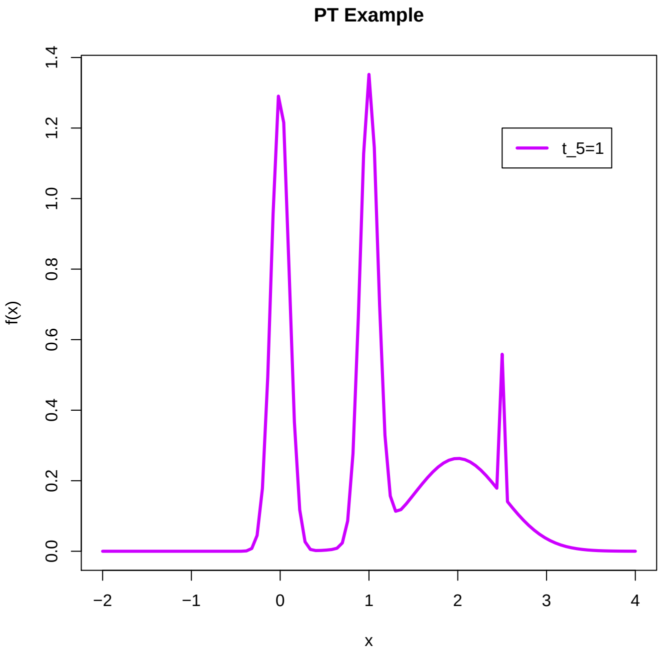
FIGURE: Full conditional log-posterior for μ_a in the AA-sampler.

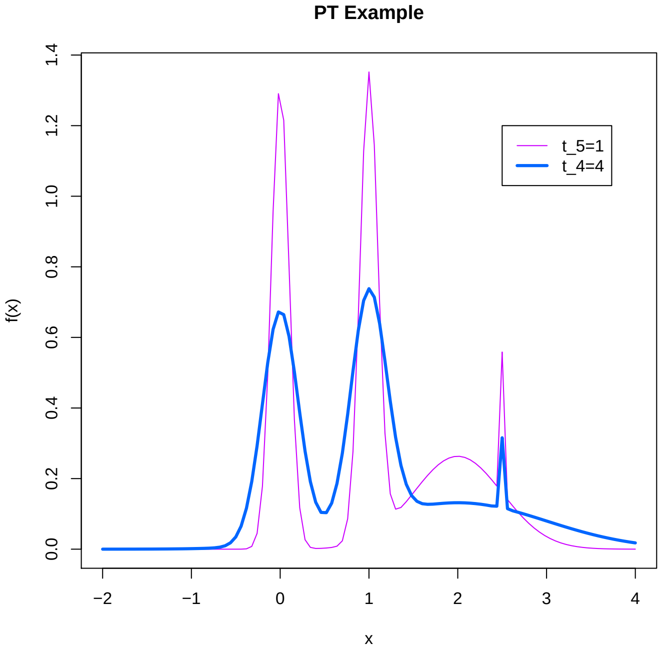
Sampling (M,A)

Sampling individual masses and ages is itself a complicated task, given the irregular and non-linear nature of the isochrone tables.

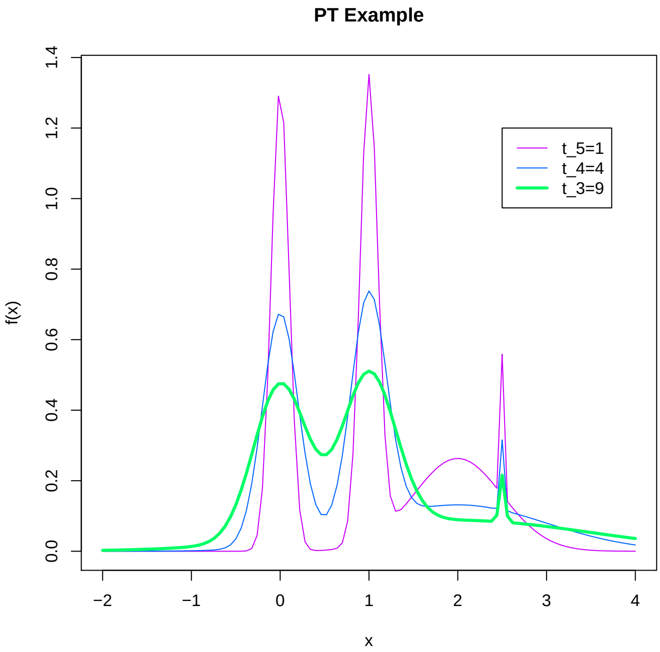
We construct an efficient, energy-based proposal distribution, in the spirit of Equi-Energy sampling (Kou, Zhou and Wong, 2006).

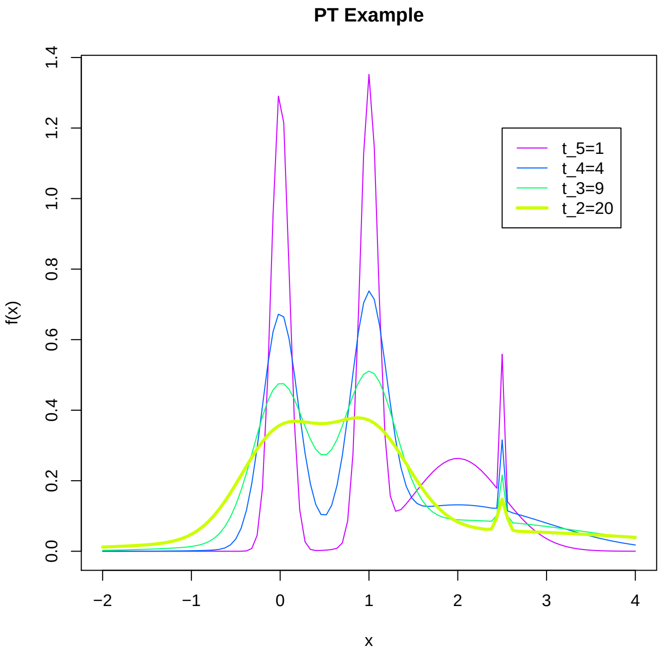
The entire scheme is then embedded within a Parallel Tempering framework (Geyer, 1991).


PT Example
t_5=1
f(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
−2
−1
0
1
2
3
4



PT Example
t_5=1
t_4=4
f(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
−2
−1
0
1
2
3
4


PT Example
t_5=1
t_4=4
t_3=9
f(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
−2
−1
0
1
2
3
4


PT Example
t_5=1
t_4=4
t_3=9
t_2=20
f(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
−2
−1
0
1
2
3
4

Parallel Tempering

Ingredients:

Parallel Tempering

Ingredients:

(1) **Energy function**: $h(x) \propto -\log \pi(x)$ (i.e. $T = 1$)

Parallel Tempering

Ingredients:

(1) **Energy function**: $h(x) \propto -\log \pi(x)$ (i.e. $T = 1$)

(2) **Temperature levels**:

$$1 = T_0 < T_1 < \cdots < T_K$$

Parallel Tempering

Ingredients:

(1) **Energy function**: $h(x) \propto -\log \pi(x)$ (i.e. $T = 1$)

(2) **Temperature levels**:

$$1 = T_0 < T_1 < \cdots < T_K$$

(3) **Tempered Targets**:

$$\pi_i(x) \propto \exp\{-h(x)/T_i\}$$

Parallel Tempering

Ingredients:

(1) **Energy function**: $h(x) \propto -\log \pi(x)$ (i.e. $T = 1$)

(2) **Temperature levels**:

$$1 = T_0 < T_1 < \cdots < T_K$$

(3) **Tempered Targets**:

$$\pi_i(x) \propto \exp\left\{-h(x)/T_i\right\}$$

(4) Hotter chains are easier to sample from (flatter), allow for a filtering of states from higher to lower temperatures (random exchange) to avoid local modes etc.

Equi-Energy Sampling

Ingredients:

Equi-Energy Sampling

Ingredients:

(1) **Energy function**: $h(x) \propto -\log \pi(x)$ (i.e. $T = 1$)

Equi-Energy Sampling

Ingredients:

(1) **Energy function**: $h(x) \propto -\log \pi(x)$ (i.e. $T = 1$)

(2) **Temperature levels**:

$$1 = T_0 < T_1 < \cdots < T_K$$

Equi-Energy Sampling

Ingredients:

(1) **Energy function**: $h(x) \propto -\log \pi(x)$ (i.e. $T = 1$)

(2) **Temperature levels**:

$$1 = T_0 < T_1 < \cdots < T_K$$

(3) **Energy levels**:

$$H_0 \leq \inf_x\{h(x)\} < H_1 < H_2 < \cdots < H_K < H_{K+1} = \infty$$

Equi-Energy Sampling

Ingredients:

(1) **Energy function**: $h(x) \propto -\log \pi(x)$ (i.e. $T = 1$)

(2) **Temperature levels**:

$$1 = T_0 < T_1 < \cdots < T_K$$

(3) **Energy levels**:

$$H_0 \leq \inf_x \{h(x)\} < H_1 < H_2 < \cdots < H_K < H_{K+1} = \infty$$

(4) **Per-level energy functions**:

$$h_i(x) = \frac{1}{T_i} \max\{h(x), H_i\} \quad \text{i.e.} \quad \pi_i(x) \propto \begin{cases} e^{-h(x)/T_i} & \text{if } h(x) > H_i \\ e^{-H_i/T_i} & \text{if } h(x) \leq H_i \end{cases}$$

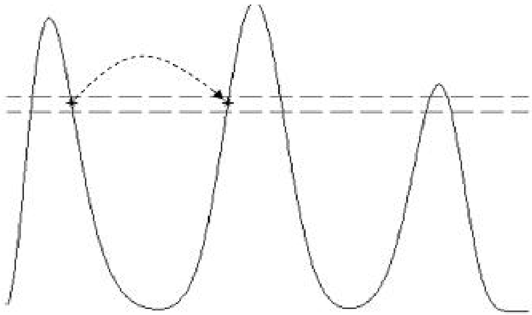


FIGURE: The equi-energy jump

Energy Levels: A Closer look

For those who are unfamiliar with temperature-scaling it may be helpful to look at exactly what we have in the familiar setting:

Energy Levels: A Closer look

For those who are unfamiliar with temperature-scaling it may be helpful to look at exactly what we have in the familiar setting:

$$\pi_i(x) = \begin{cases} (\pi(x))^{1/T_i} & \text{if } \pi(x) > e^{-H_i} \\ \text{constant} & \text{if } \pi(x) \leq e^{-H_i} \end{cases}$$

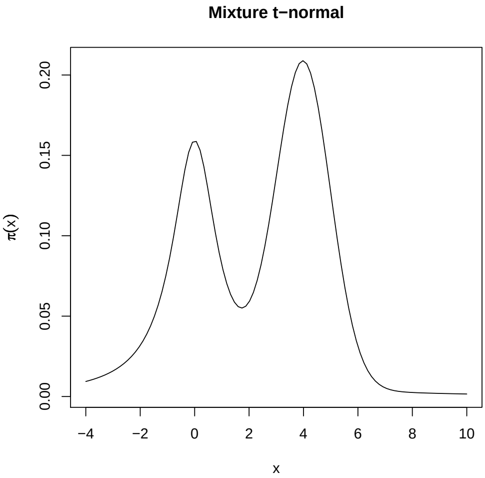
Energy Levels: A Closer look

For those who are unfamiliar with temperature-scaling it may be helpful to look at exactly what we have in the familiar setting:

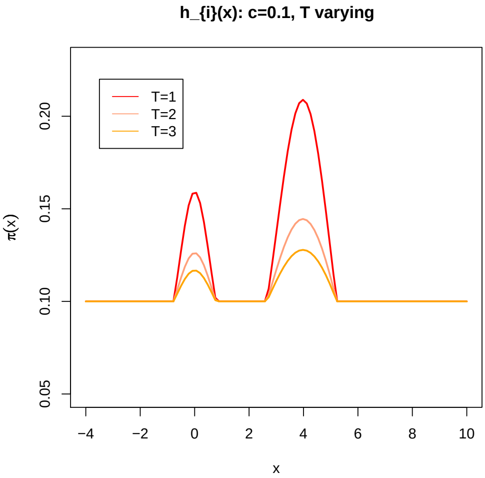
$$\pi_i(x) = \begin{cases} (\pi(x))^{1/T_i} & \text{if } \pi(x) > e^{-H_i} \\ \text{constant} & \text{if } \pi(x) \leq e^{-H_i} \end{cases}$$

- So the $i^{th}-$level target densities are **(lower-)truncated and flattened** versions of the original density

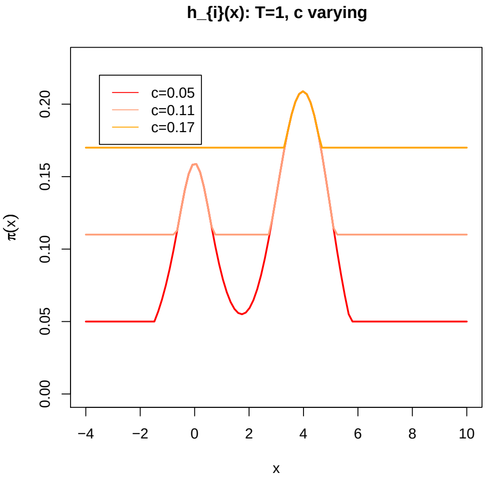
An example



Temperature Effect



Truncation Effect



THE DETAILS

- **Idea:** Construct empirical energy rings $\hat{D}_j^{(K)}$ for $j = 0, 1, \ldots, K$
  i.e. compute set membership $I(x)$ for each sampled state:

$$I(x) = j \quad \text{if} \quad h(x) \in [H_j, H_{j+1}) \quad \left(\text{ i.e. } \pi(x) \in \ [e^{-H_j}, e^{-H_{j+1}}) \right)$$

EE-inspired Proposal Construction

The driving force of the posterior in the ancillary augmentation is the likelihood, as determined primarily by the isochrone tables.

- EE constructs energy bins in joint space...

EE-inspired Proposal Construction

The driving force of the posterior in the ancillary augmentation is the likelihood, as determined primarily by the isochrone tables.

- EE constructs energy bins in joint space...
  here $2n + 2 + p(p-1)/2$

EE-inspired Proposal Construction

The driving force of the posterior in the ancillary augmentation is the likelihood, as determined primarily by the isochrone tables.

- EE constructs energy bins in joint space...
  here $2n + 2 + p(p-1)/2$ ☹

EE-inspired Proposal Construction

The driving force of the posterior in the ancillary augmentation is the likelihood, as determined primarily by the isochrone tables.

- EE constructs energy bins in joint space...
  here $2n + 2 + p(p-1)/2$ ☹
- For Gibbs step drawing (M_i, A_i), we can utilize the common structure of the isochrone tables
- We have a map ahead of time of which parameter combinations map to 'similar' regions of the observation space
- Construct 'energy-based' proposal based on this
- Partition parameter space into 'boxes' and characterize distance between boxes...

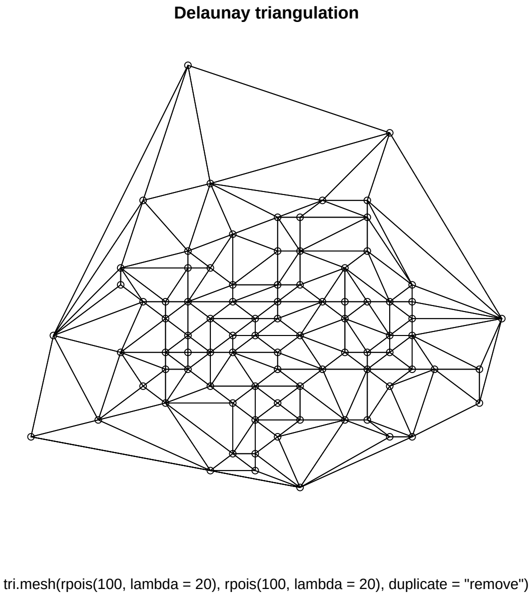


FIGURE: Example of Delaunay Triangulation

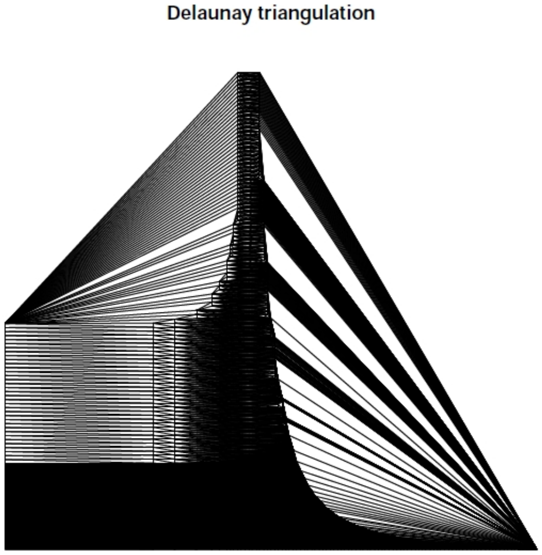


FIGURE: Delaunay Triangulation of (M, A)

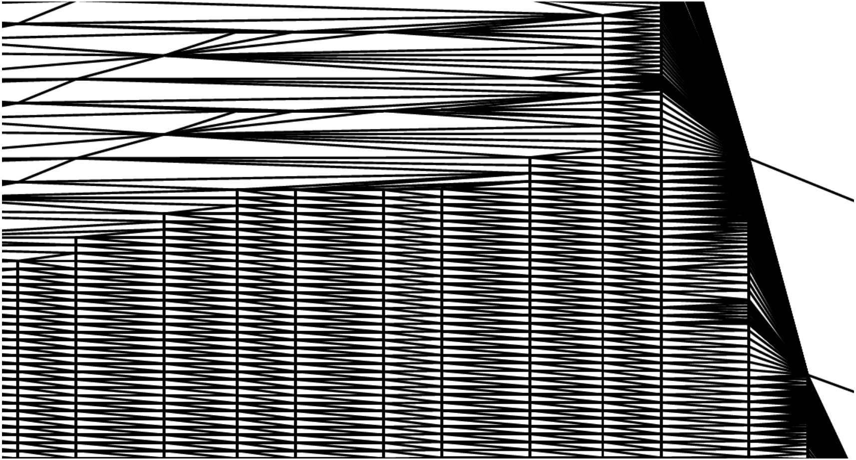


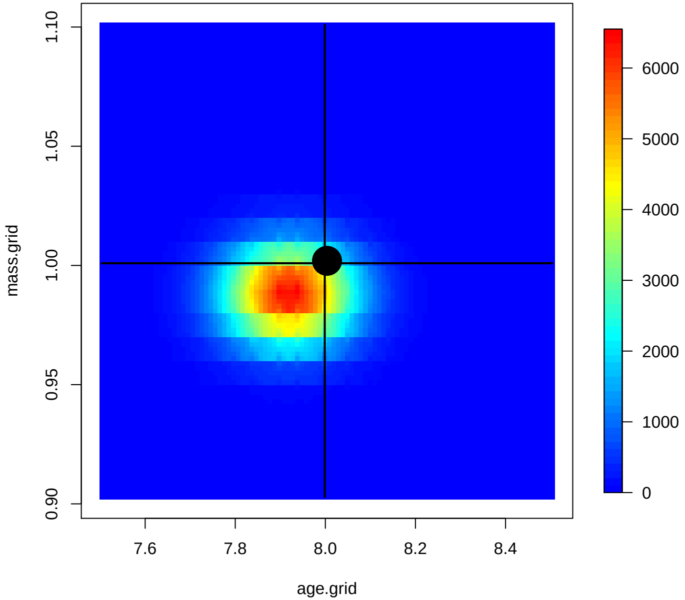
FIGURE: Close up view of the partition

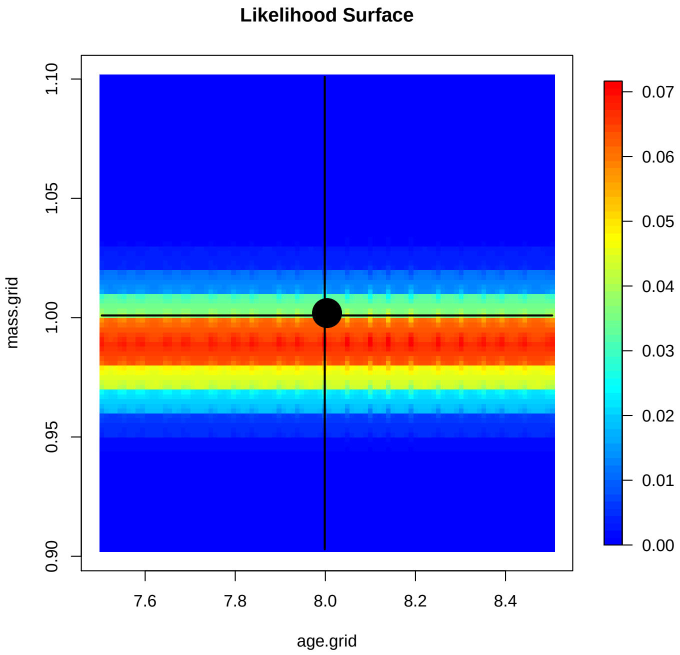
Parsimonious Representation

Storging a ‘box-to-box’ proposal weight matrix is a little cumbersome as the number of boxes is relatively large ($\approx 193,000$).

Instead we cluster the boxes into ‘clusters’ of boxes that are ‘close’ in color/magnitude space, then propose (uniformly) within the clusters, and store a much smaller cluster-to-cluster proposal.

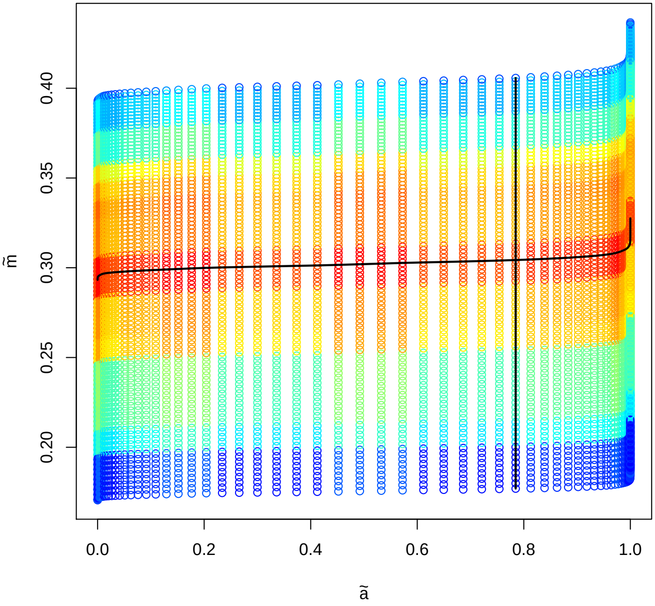
Posterior distribution



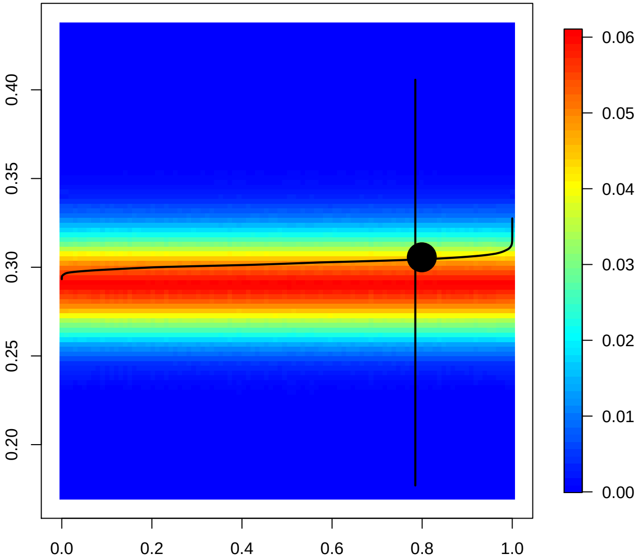


Likelihood Surface
mass.grid
age.grid
1.10
1.05
1.00
0.95
0.90
7.6
7.8
8.0
8.2
8.4
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00

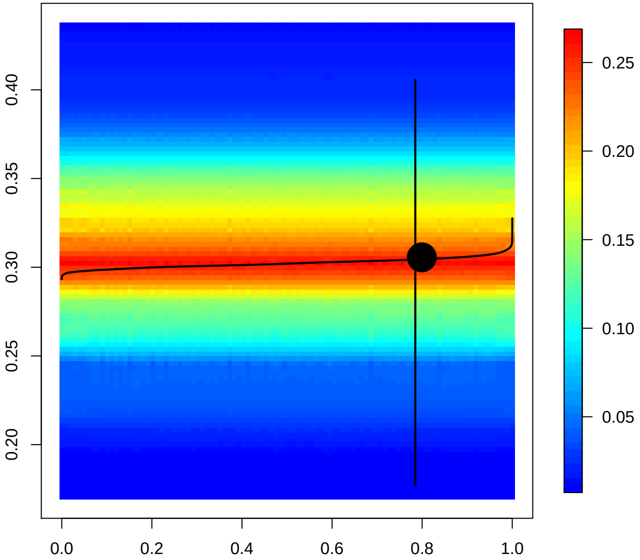
Transformation: SA−>AA (transition)



Posterior in AA−parametrization



Proposal distribution in AA−parametrization



Implementation

The goal was a general purpose methodology for CMD-esque analysis, for which we have developed (and are developing) an `R` package `cmd`, complete with other computational and graphical tools for CMD-analysis.

The algorithm is implemented in `C`, and can effectively datasets up to $\approx 2,000,000$ observations.

The package can be run in standard form or, for more intensive analyses, a multi-threaded implementation is available for parallel computation.

Python front end expected in the future...

Results

Simulating from the model and fitting, we check the sampling scheme validates (on average)...

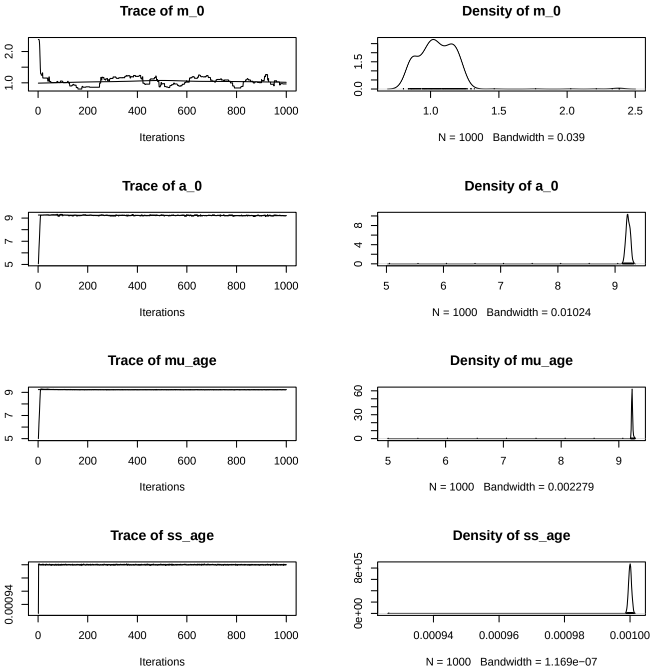
Results

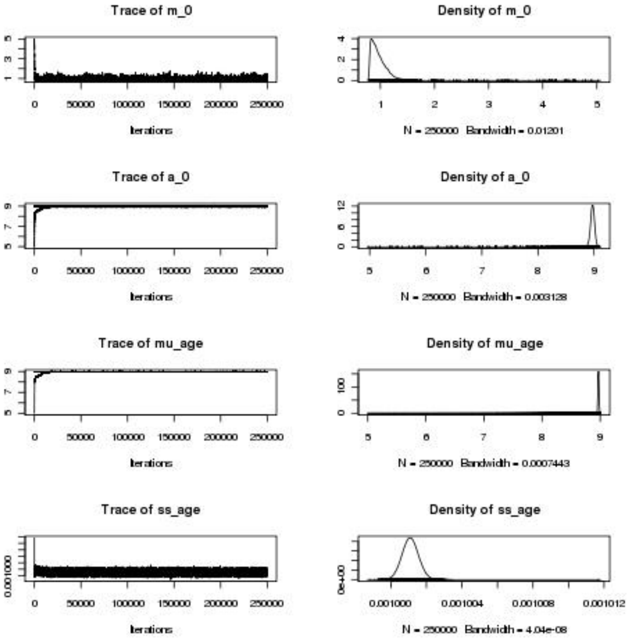
Simulating from the model and fitting, we check the sampling scheme validates (on average)...

Validation Recipe (Aggregate)

1. *Simulate parameters from the prior distribution, M times,*
2. *For each parameter combination, simulate a dataset conditional on these,*
3. *Compute the posterior interval(s) for each of these datasets,*
4. *Compute whether or not each interval covers the truth.*

Coverage rates of 95% interval (etc) should agree with nominal level.


Trace of m_0
Iterations
Density of m_0
N = 1000   Bandwidth = 0.039
Trace of a_0
Iterations
Density of a_0
N = 1000   Bandwidth = 0.01024
Trace of mu_age
Iterations
Density of mu_age
N = 1000   Bandwidth = 0.002279
Trace of ss_age
Iterations
Density of ss_age
N = 1000   Bandwidth = 1.169e−07



Trace of m_0
Iterations
Density of m_0
N = 250000 Bandwidth = 0.01201
Trace of a_0
Iterations
Density of a_0
N = 250000 Bandwidth = 0.003128
Trace of mu_age
Iterations
Density of mu_age
N = 250000 Bandwidth = 0.0007443
Trace of ss_age
Iterations
Density of ss_age
N = 250000 Bandwidth = 4.04e-08

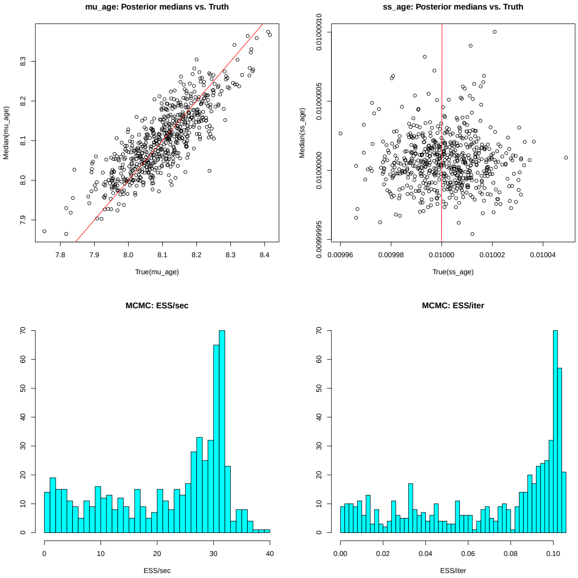


FIGURE: Convergence plots for sampler with Parallel Tempering

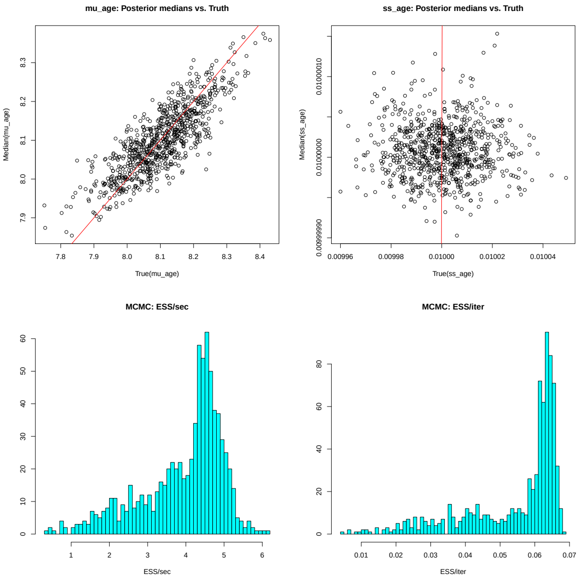
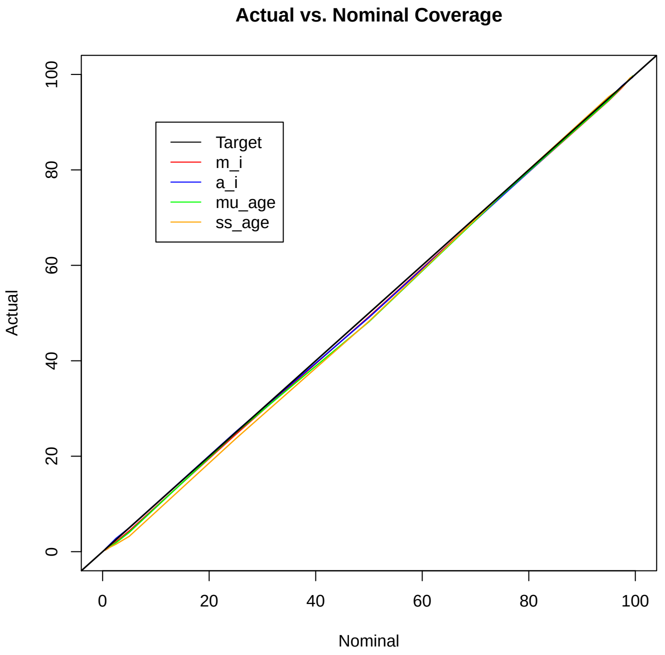


FIGURE: Convergence plots for sampler with box-and-cluster proposal in the ancillary augmentation


Actual vs. Nominal Coverage
Target
m_i
a_i
mu_age
ss_age
Actual
Nominal
0
20
40
60
80
100

Coverage Table

	1%	2.5%	5%	25%	50%	75%	95%	97.5%	99%
m_i	0.7	2.1	4.4	24.6	49.3	74.6	94.5	97.2	99.2
a_i	1.1	2.8	5.0	25.1	49.1	74.5	94.8	97.6	99.0
mu_a	0.7	1.9	4.0	24.9	48.2	74.8	94.6	97.4	99.1
sigmasq_a	0.7	1.6	3.2	23.7	48.3	75.0	95.3	97.3	99.3

Table: Coverage table for the standard SA+AA box+cluster sampler with Geneva isochrones.

Summary

- Use a flexible hierarchical Bayesian model to make inference about properties of stellar clusters
- The complex 'black-box' nature of the lookup table presents computation challenges
- By interweaving two complementary parametrizations we can gain relative to each individual scheme
- Distinction between inferential and computational parametrizations
- Combination of several heavy-hitting MCMC techinques allows us to effectively sample from the posterior for wide range of lookup tables

References

Acknowledgments

1. Kou, S.C., Zhou, Q., Wong, W.H., (2006) Equi-Energy Sampler with applications in statistical inference and statistical mechanics. *The Annals of Statistics*, **34**-4, pp.1581-1619
2. Geyer, C.J. (1991) Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics Proceedings of the 23rd Symposium on the Interface*, p. 156-163.
3. Baines, P.D., Meng, X.-L., Zezas, A., Kashyap, V.K., Lee, H. (2009) Bayesian Analysis of Color-Magnitude Diagrams (in progress)
4. Yu, Y., Meng, X.-L. (2009) To Center or Not To Center: That Is Not The Question, (in progress)