Expectation Maximization - Missing Data

Matt Oehler

Brigham Young University

Stat 624 Project 2

December 11, 2017

Overview

Introduction/Motivation

Problem

Parameter estimation is important, and there are various methods used to solve these kinds of problems.

Dilemma

Many approaches are not robust when they encounter non-ideal circumstances (e.g. missing data).

Solution

It is possible to work around this dilemma through various data imputation methods, such as the Expectation-Maximization Algorithm

Questions of Interest

Question 1

Can we still find good estimates of the means and covariances between variables in the presence of missing data?

Question 2

Can we come up with a method to determine if the data are missing at random or if there is a pattern to the missingness?

Methodology

In this study we will look at 3 different methods of data imputation for data that follow a multivariate normal distribution.
Methods:

- Throw-away Method
- Expectation-Maximization Algorithm
- Conditional Sampling

MVN PDF:

$$f(x|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} exp\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\}$$

Methodology

'Throw-away' Method:

- This method isn't actually an imputation method. It simply entails the removal of all non-complete observations, and then using the sample mean and sample covariance as parameter estimates.

Mean Estimate (MLE)

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N}(x_n)$$

Covariance Estimate (Unbiased)

$$\hat{\Sigma} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^\top$$

Methodology

Expectation-Maximization Algorithm:

- Instead of throwing away the incomplete data, we can iteratively update estimates for the mean. The algorithm will continue to iterate through 'expectation' and 'maximization' steps until it converges. The updates are made using the condition multivariate normal distribution.

$$(\mathbf{x}_1|\mathbf{x}_2 = a) \sim N(\bar{\mu}, \bar{\Sigma})$$

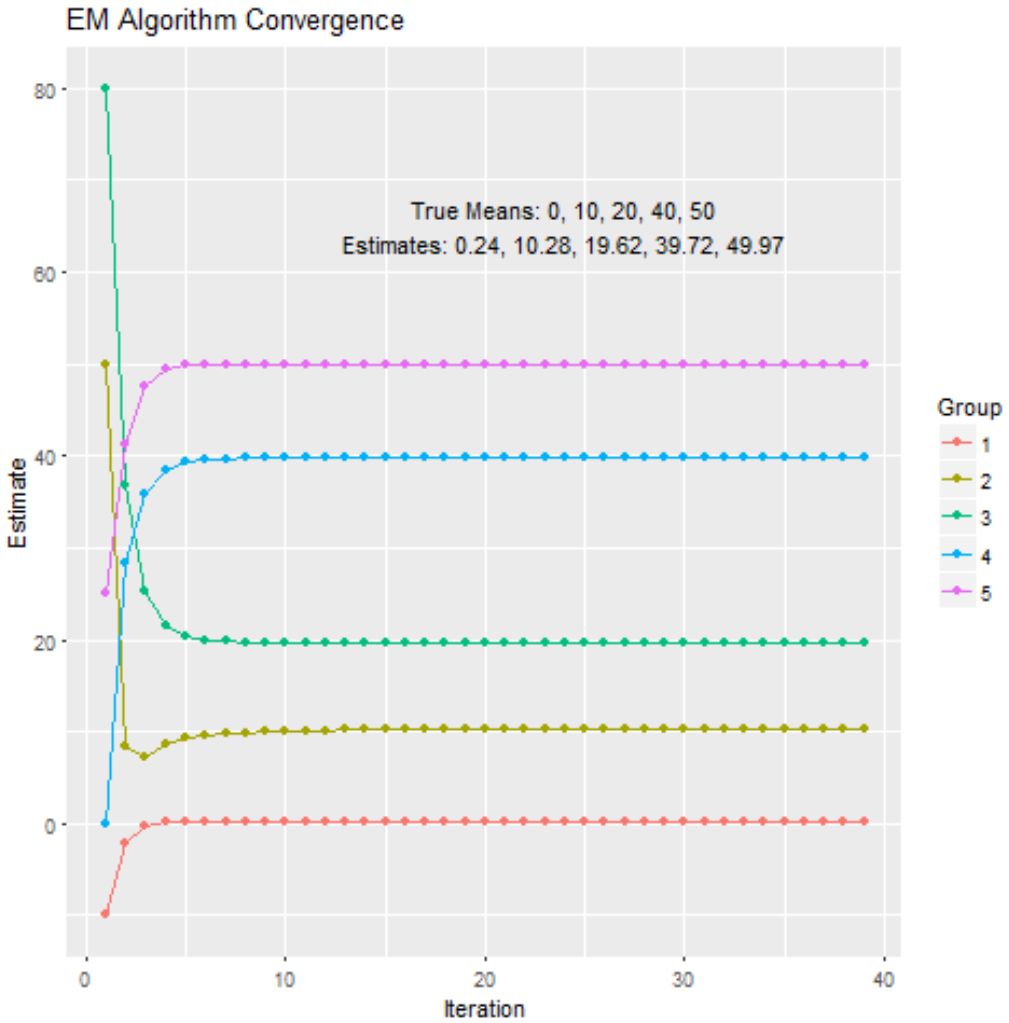
$$\bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$$

$$\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Note: subscript 1 refers to missing values, and subscript 2 refers to non-missing values

Methodology

Convergence Plot:



Methodology

'Conditional Sampling':

- Similar to the EM Algorithm, but instead of imputing the mean for the missing values, we draw random values from the distribution of the estimates of μ and Σ for each iteration. The draws of values are then used to estimate the mean and covariance. (This method will not converge)

$$(\mathbf{x}_1|\mathbf{x}_2 = a) \sim N(\bar{\mu}, \bar{\Sigma})$$

$$\bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$$

$$\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Note: subscript 1 refers to missing values, and subscript 2 refers to non-missing values

MC Simulated Means for randomly missing data:

Method	Estimate	Bias	MSE
Throw Away	0.0010	0.0010	0.0621
	10.0021	0.0021	0.0612
	20.0023	0.0023	0.0615
	40.0026	0.0026	0.0615
	50.0034	0.0034	0.0620
EM Algorithm	0.0008	0.0008	0.0124
	10.0049	0.0049	0.0135
	20.0048	0.0048	0.0139
	40.0022	0.0022	0.0125
	50.0027	0.0027	0.0144
Conditional Sampling	-0.0062	-0.0062	0.0122
	9.9961	-0.0039	0.0126
	19.9973	-0.0027	0.0142
	39.9920	-0.0080	0.0118
	49.9990	-0.0010	0.0157

MC Simulated Data for non-random missing data:

	Estimate	Bias	MSE
Throw Away	-0.2541	-0.2541	0.0725
	9.7889	-0.2111	0.0527
	19.7803	-0.2197	0.0566
	39.7524	-0.2476	0.0689
	49.7926	-0.2074	0.0514
EM Algorithm	-0.0931	-0.0931	0.0172
	9.8625	-0.1375	0.0234
	19.8898	-0.1102	0.0162
	39.8926	-0.1074	0.0154
	49.8481	-0.1519	0.0307
Conditional Sampling	-0.0871	-0.0871	0.0155
	9.9055	-0.0945	0.0165
	19.8943	-0.1057	0.0188
	39.9142	-0.0858	0.0150
	49.8757	-0.1243	0.0229

Application

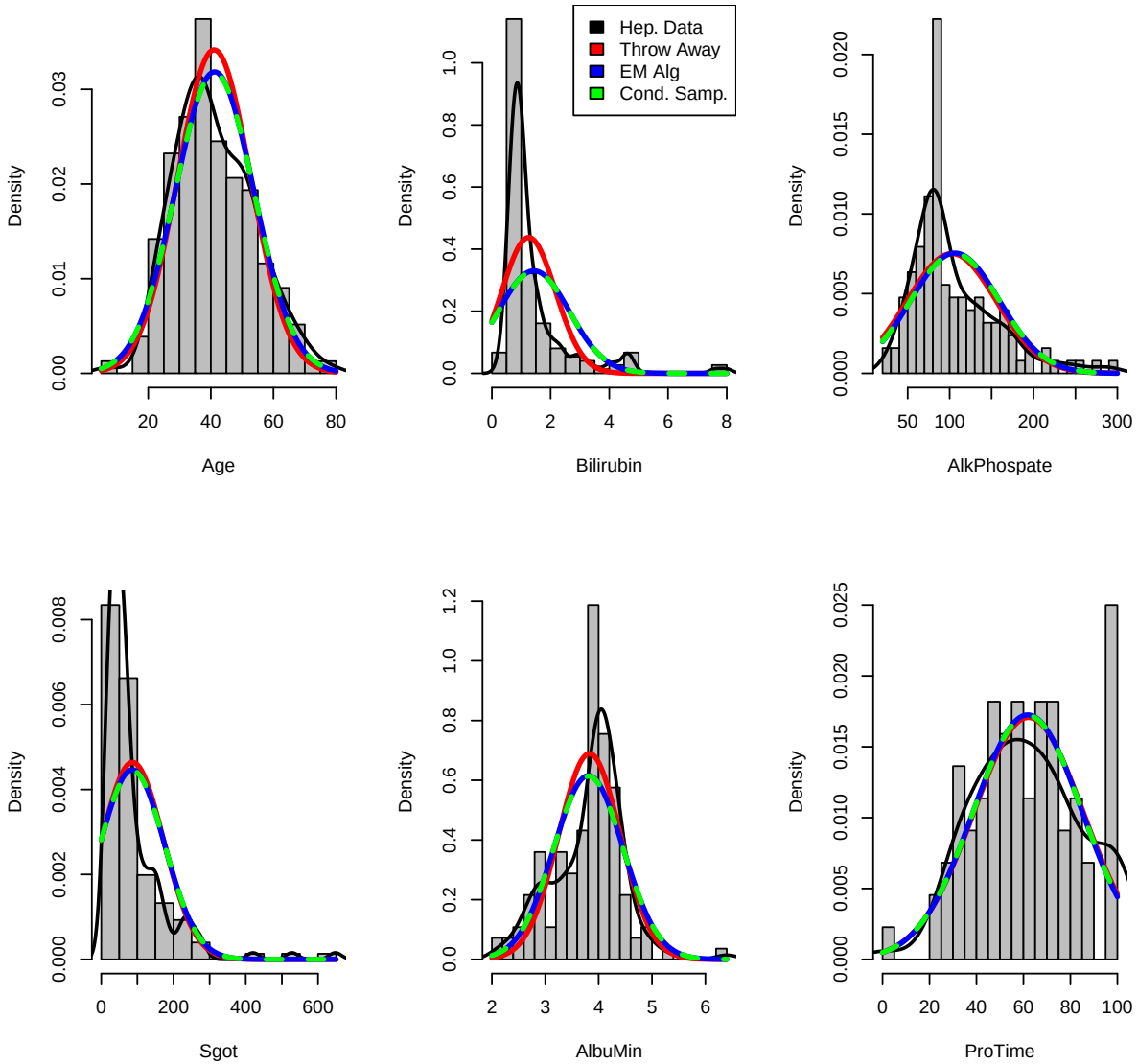
We'll test these methods using data of characteristics of hepatitis patients:

Age	Bilirubin	AlkPhosphate	Sgot	AlbuMin	ProTime
30	1.00	85	18	4.00	
50	0.90	135	42	3.50	
78	0.70	96	32	4.00	
31	0.70	46	52	4.00	80
34	1.00		200	4.00	
34	0.90	95	28	4.00	75

Application

	Throw Away	EM Algorithm	Conditional Sampling
Age	41.06	41.20	41.20
Bilirubin	1.25	1.43	1.43
AlkPhosphate	102.51	106.30	106.31
Sgot	86.39	85.89	85.92
AlbuMin	3.83	3.81	3.81
ProTime	62.16	61.81	61.74

Goodness of Fit



Conclusions

- Even with missing data, we can find good parameter estimates
- The Expectation-Maximization Algorithm and Conditional Sampling perform very similarly
- We can get an idea of how (random or non-random) data are missing by comparing the results of the various methods

The End