1. Introduction: The research question should be introduced with some motivations as to why that particular question may be important. The methodology and data used is also introduced along with some other potential areas where the methodology is useful.

2. Methodology: The methodology is described in detail. How do you plan to answer the research questions? What other statistical techniques are involved (i.e. optimization, regression, hypothesis testing)?

3. Simulation Study: Prove the methodology you proposed in the previous section works the way you say it will work to answer the question.

4. Results: Introduce and explore the data. Apply the methodology to the data and report/display your results.

5. Conclusion: A short rehash of the main ideas from the previous four sections.

Intro:

Explain why missing data can be an issue, and how there are multiple approaches to fixing it.

Include some history of the EM algorithm.

* Nature.com page had some stuff

Explain the dataset. And why I will be using it

Estimating the parameters of a multivariate normal distribution is relatively simple. However, when the data set isn’t complete (i.e. there are missing values) then parameter estimation significantly increases in complexity. There are several approaches that can be taken when values are missing from a dataset for which one want to estimate the distribution parameters. In this study, we will look at 3 separate approaches, and compare their results. The first method is to simply remove all incomplete observations from the dataset and then estimate parameter values with the remaining observations. The second method is known as Expectation Maximization, or the EM algorithm, and the third method is conditional normal theory (estimating parameters based on the conditional multivariate normal distribution). Each method has its own respective pros and cons, which will be discussed in further detail throughout the paper. After comparing the results of each method, we will look into finding a pattern in the missing data, finding mean and covariance estimates between the variables, and determine which variables are most and least correlated. To compare these methods, we will use data of characteristics of individuals with hepatitis. The data set contains several characteristics, but we will only be using the non-categorical variables, which are: Age, Bibirubin, alkphosphate, sgot alumin, and protime. (RELIST THESE VARIABLES AS WHAT THEY ACTUALLY MEAN). The data includes observations of people with and without hepatitis, but the focus of this study will not be looking at ways to better predict patients that have the disease or not. Instead we will just be focusing on methods of estimating the distribution parameters in the presence of missing data.

Methodology

Show the results of parameter estimates after removing all observations that have missing values

Use expectation maximization

Conditional normal theory