Soc504: Mixtures, EM and Missing Data

Brandon Stewart¹

Princeton

March 27- April 5, 2017

¹The EM section draws on some slides from Justin Grimmer, Patrick Lam and generations of teaching assistants for Gov2001 at Harvard. The missing data section draws heavily on slides from Gary King. The measurement error section draws heavily on slides from Matt Blackwell.
Readings

- Monday (Mixture Models)


- Wednesday (EM)


- Monday (Missing Data)


- Wednesday (Missing Data)

Readings

- **Monday (Mixture Models)**

- **Wednesday (EM)**

- **Monday (Missing Data)**

- **Wednesday (Missing Data)**
Readings

Monday (Mixture Models)

Wednesday (EM)

Monday (Missing Data)

Wednesday (Missing Data)
Readings

**Monday (Mixture Models)**


**Wednesday (EM)**


**Monday (Missing Data)**

Readings

- **Monday (Mixture Models)**

- **Wednesday (EM)**

- **Monday (Missing Data)**

- **Wednesday (Missing Data)**
Readings

Monday (Mixture Models)

Wednesday (EM)
Readings

- **Monday (Mixture Models)**

- **Wednesday (EM)**

- **Monday (Missing Data)**

- **Stewart (Princeton)**
  - Missing Data
  - Mar 27-Apr 5, 2017 2 / 200
Readings

Monday (Mixture Models)

Wednesday (EM)

Monday (Missing Data)
Readings

- **Monday (Mixture Models)**

- **Wednesday (EM)**

- **Monday (Missing Data)**
Readings

Monday (Mixture Models)

Wednesday (EM)

Monday (Missing Data)

Wednesday (Missing Data)
Readings

Monday (Mixture Models)

Wednesday (EM)

Monday (Missing Data)

Wednesday (Missing Data)
1 Mixture Models
- Basic Mixtures
- Application: Mixtures as Preprocessing
- Application: Mixture of Regressions

2 Expectation Maximization
- EM for Probit Regression
- EM for Gaussian Mixtures
- EM in General

3 Missing Data
- Motivating Example
- Overview and Assumptions
- Existing Heuristics
- Application Specific Approaches
- Multiple Imputation
- The Full Amelia Scheme

4 Measurement Error

5 Appendix: Additional Details and Examples
# Old Faithful

## Old Faithful Eruption Times

<table>
<thead>
<tr>
<th>Eruption Length</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>0</td>
</tr>
<tr>
<td>2.0</td>
<td>5</td>
</tr>
<tr>
<td>2.5</td>
<td>10</td>
</tr>
<tr>
<td>3.0</td>
<td>15</td>
</tr>
<tr>
<td>3.5</td>
<td>20</td>
</tr>
<tr>
<td>4.0</td>
<td>0</td>
</tr>
<tr>
<td>4.5</td>
<td>10</td>
</tr>
<tr>
<td>5.0</td>
<td>5</td>
</tr>
</tbody>
</table>

Stewart (Princeton)  
Mar 27-Apr 5, 2017 4 / 200
Old Faithful

How do we summarize? No handy distribution.

We can try fitting a normal but the fit is poor.

If you squint, it looks like two different normals.

Stewart (Princeton)
How do we summarize? No handy distribution
How do we summarize? No handy distribution

We can try fitting a normal but the fit is poor
Old Faithful

How do we summarize? No handy distribution
We can try fitting a normal but the fit is poor
If you squint, it looks like two different normals
Mixture Models

- Mixtures allow us to represent a more complex data generating process.
Mixture Models

- Mixtures allow us to represent a more complex data generating process
- Working backwards, we want two normal distributions. Let’s introduce $z_i \in \{1, 2\}$ to indicate which normal distribution observation $i$ comes from.

When $z_i = 1$ we see $p(y_i | z_i = 1) \sim N(\mu_1, \sigma^2_1)$

When $z_i = 2$ we see $p(y_i | z_i = 2) \sim N(\mu_2, \sigma^2_2)$

To complete the model we give $z_i$ a distribution $z_i \sim \text{Bernoulli}(\pi)$

Our goal is to estimate $\mu_1, \mu_2, \sigma^2_1, \sigma^2_2, \pi$ However, we don’t observe $z_i$, this is a type of missing data.
Mixture Models

- Mixtures allow us to represent a more complex data generating process
- Working backwards, we want two normal distributions. Let’s introduce \( z_i \in \{1, 2\} \) to indicate which normal distribution observation \( i \) comes from.
- When \( z_i = 1 \) we see \( p(y_i|z_i = 1) \sim \mathcal{N}(\mu_1, \sigma_1^2) \)
Mixture Models

- Mixtures allow us to represent a more complex data generating process.
- Working backwards, we want two normal distributions. Let’s introduce \( z_i \in \{1, 2\} \) to indicate which normal distribution observation \( i \) comes from.
- When \( z_i = 1 \) we see \( p(y_i | z_i = 1) \sim \mathcal{N}(\mu_1, \sigma_1^2) \)
- When \( z_i = 2 \) we see \( p(y_i | z_i = 2) \sim \mathcal{N}(\mu_2, \sigma_2^2) \)
Mixture Models

- Mixtures allow us to represent a more complex data generating process.
- Working backwards, we want two normal distributions. Let’s introduce $z_i \in \{1, 2\}$ to indicate which normal distribution observation $i$ comes from.
- When $z_i = 1$ we see $p(y_i|z_i = 1) \sim \mathcal{N}(\mu_1, \sigma_1^2)$
- When $z_i = 2$ we see $p(y_i|z_i = 2) \sim \mathcal{N}(\mu_2, \sigma_2^2)$
- To complete the model we give $z_i$ a distribution $z_i \sim \text{Bernoulli}(\pi)$
Mixture Models

- Mixtures allow us to represent a more complex data generating process.
- Working backwards, we want two normal distributions. Let’s introduce $z_i \in \{1, 2\}$ to indicate which normal distribution observation $i$ comes from.
- When $z_i = 1$ we see $p(y_i|z_i = 1) \sim \mathcal{N}(\mu_1, \sigma_1^2)$
- When $z_i = 2$ we see $p(y_i|z_i = 2) \sim \mathcal{N}(\mu_2, \sigma_2^2)$
- To complete the model we give $z_i$ a distribution $z_i \sim \text{Bernoulli}(\pi)$
- Our goal is to estimate $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi$
Mixture Models

- Mixtures allow us to represent a more complex data generating process.
- Working backwards, we want two normal distributions. Let’s introduce $z_i \in \{1, 2\}$ to indicate which normal distribution observation $i$ comes from.
- When $z_i = 1$ we see $p(y_i|z_i = 1) \sim \mathcal{N}(\mu_1, \sigma_1^2)$
- When $z_i = 2$ we see $p(y_i|z_i = 2) \sim \mathcal{N}(\mu_2, \sigma_2^2)$
- To complete the model we give $z_i$ a distribution $z_i \sim \text{Bernoulli}(\pi)$
- Our goal is to estimate $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi$
- However, we don’t observe $z_i$, this is a type of missing data.
Mixture Models

- Notice that we don’t need the $z$ variable indicators at all, they are just a convenience for specifying the model.
Mixture Models

- Notice that we don’t need the $z$ variable indicators at all, they are just a convenience for specifying the model.
- We can write the log likelihood as

$$\ell = \sum_i \left( \log \left( \sum_k \pi_k \mathcal{N}(x_i | \mu_k, \sigma_k^2) \right) \right)$$
Mixture Models

- Notice that we don’t need the $z$ variable indicators at all, they are just a convenience for specifying the model.
- We can write the log likelihood as

$$
\ell = \sum_i \left( \log \left( \sum_k \pi_k \mathcal{N}(x_i | \mu_k, \sigma_k^2) \right) \right)
$$

- This is hard to solve due to the summation inside the log.
Mixture Models

- Notice that we don’t need the $z$ variable indicators at all, they are just a convenience for specifying the model.
- We can write the log likelihood as

$$\ell = \sum_i \left( \log \left( \sum_k \pi_k \mathcal{N}(x_i | \mu_k, \sigma_k^2) \right) \right)$$

- This is hard to solve due to the summation inside the log.
- By introducing the missing variables $z$, we make it easier to estimate the parameters. This is called data augmentation.
Mixture Models

- Notice that we don’t need the $z$ variable indicators at all, they are just a convenience for specifying the model.
- We can write the log likelihood as

$$\ell = \sum_i \left( \log \left( \sum_k \pi_k \mathcal{N}(x_i | \mu_k, \sigma_k^2) \right) \right)$$

- This is hard to solve due to the summation inside the log.
- By introducing the missing variables $z$, we make it easier to estimate the parameters. This is called data augmentation.
- This problem was easy because the components are well separated.
A Harder Problem

Student Heights

Some distributions have less clear separation
A Harder Problem

Bimodality here arises due to gender
The mixture model *sort of* captures this
A Harder Problem

Height by Sex

Density

Height in cm

The true distributions are more peaked with fatter tails
One component captures all the women but also many men.
Multiple Dimensions

This strategy also works in more than one dimension.

Now the cluster indicator indexes a multivariate distribution.

This fits the data reasonably well.
Multiple Dimensions

This strategy also works in more than one dimension
This strategy also works in more than one dimension

Now the cluster indicator indexes a multivariate distribution
Multiple Dimensions

- This strategy also works in more than one dimension
- Now the cluster indicator indexes a multivariate distribution
- This fits the data reasonable well
The Gist of Computation

From Bishop (2006) Chapter 9
The Gist of Computation

$L = 1$

From Bishop (2006) Chapter 9
The Gist of Computation

From Bishop (2006) Chapter 9

(d) $L = 2$
The Gist of Computation

From Bishop (2006) Chapter 9
The Gist of Computation

$L = 20$

From Bishop (2006) Chapter 9
Mixture Models Can Have Many Components
Mixture Models Can Have Many Components

Imagine we draw from data with a 3 component mixture.
Mixture Models Can Have Many Components

We observe only the data without the labels.
Mixture Models Can Have Many Components

But we can still infer the components well
Basic Mixtures

- Simple mixture models can be a useful way to model complicated distributions
Basic Mixtures

- Simple mixture models can be a useful way to model complicated distributions.
- We saw a heuristic version of the computation, it is a form of a general algorithm called *Expectation Maximization (EM)* which will be useful in many contexts.
Basic Mixtures

- Simple mixture models can be a useful way to model complicated distributions.
- We saw a heuristic version of the computation, it is a form of a general algorithm called Expectation Maximization (EM) which will be useful in many contexts.
- Estimation leverages the idea of data augmentation which also shows up in a number of areas of computational statistics.
Basic Mixtures

- Simple mixture models can be a useful way to model complicated distributions.
- We saw a heuristic version of the computation, it is a form of a general algorithm called **Expectation Maximization (EM)** which will be useful in many contexts.
- Estimation leverages the idea of **data augmentation** which also shows up in a number of areas of computational statistics.
- The mixture model framework can also be used in various other models.
Basic Mixtures

- Simple mixture models can be a useful way to model complicated distributions.
- We saw a heuristic version of the computation, it is a form of a general algorithm called Expectation Maximization (EM) which will be useful in many contexts.
- Estimation leverages the idea of data augmentation which also shows up in a number of areas of computational statistics.
- The mixture model framework can also be used in various other models.
- For example, Latent Class Analysis is a mixture of multinomials model commonly used to analyze surveys.
Two Applications

Two Applications

Two Applications

- Two articles motivated from a common methodological place
Two Applications

- Two articles motivated from a common methodological place
- Both use mixtures in the context of regression
Massey et al (1993, 1994, 1998) argue that multiple mechanisms drive migrants. There can be migrants who are income-maximizing and those attracted by family. These are not mutually exclusive.
Multiple Mechanisms of Migration

- Massey et al (1993, 1994, 1998) argue that multiple mechanisms drive migrants. There can be migrants who are income-maximizing and those attracted by family. These are not mutually exclusive.
- Massey and Espinosa (1997) test in the Mexico-U.S. setting by putting a bunch of variables in a regression to see which best predicted the outcome.
Multiple Mechanisms of Migration

- Massey et al (1993, 1994, 1998) argue that multiple mechanisms drive migrants. There can be migrants who are income-maximizing and those attracted by family. These are not mutually exclusive.

- Massey and Espinosa (1997) test in the Mexico-U.S. setting by putting a bunch of variables in a regression to see which best predicted the outcome.

- This treats explanations as competing and thus is inconsistent with the theory (which says different migrants can be motivated by different things).
Multiple Mechanisms of Migration

- Massey et al (1993, 1994, 1998) argue that multiple mechanisms drive migrants. There can be migrants who are income-maximizing and those attracted by family. These are not mutually exclusive.
- Massey and Espinosa (1997) test in the Mexico-U.S. setting by putting a bunch of variables in a regression to see which best predicted the outcome.
- This treats explanations as competing and thus is inconsistent with the theory (which says different migrants can be motivated by different things).
- Instead we would prefer to identify the unknown groups of migrants who are best explained by each theory.
Multiple Mechanisms of Migration

- Massey et al (1993, 1994, 1998) argue that multiple mechanisms drive migrants. There can be migrants who are income-maximizing and those attracted by family. These are not mutually exclusive.
- Massey and Espinosa (1997) test in the Mexico-U.S. setting by putting a bunch of variables in a regression to see which best predicted the outcome.
- This treats explanations as competing and thus is inconsistent with the theory (which says different migrants can be motivated by different things).
- Instead we would prefer to identify the unknown groups of migrants who are best explained by each theory.
- We are interested in heterogeneity which is masked by missing groups.
Garip (2012) uses the Mexican Migration Project data and discovers four types of migrants.
Garip (2012)

- Garip (2012) uses the Mexican Migration Project data and discovers four types of migrants.
- The approach is algorithmic rather than probabilistic, i.e. the task is framed as an optimization problem rather than a data generating process.
Garip (2012) uses the Mexican Migration Project data and discovers four types of migrants.

The approach is algorithmic rather than probabilistic, i.e. the task is framed as an optimization problem rather than a data generating process.

Each type of identified migrant has a distinct configuration of individual, household and community characteristics.
Garip (2012) uses the Mexican Migration Project data and discovers four types of migrants.

The approach is algorithmic rather than probabilistic, i.e. the task is framed as an optimization problem rather than a data generating process.

Each type of identified migrant has a distinct configuration of individual, household and community characteristics.

Garip (2012) outlines the steps in cluster analysis as choosing:

1. the relevant attributes
2. an algorithm
3. a similarity measure
4. number of clusters or mixture components
5. validation strategy

After dividing the units, separate regressions are estimated for each cluster.
Garip (2012) uses the Mexican Migration Project data and discovers four types of migrants.

The approach is algorithmic rather than probabilistic, i.e. the task is framed as an optimization problem rather than a data generating process.

Each type of identified migrant has a distinct configuration of individual, household and community characteristics.

Garip (2012) outlines the steps in cluster analysis as choosing:

1. the relevant attributes
Garip (2012) uses the Mexican Migration Project data and discovers four types of migrants.

The approach is algorithmic rather than probabilistic, i.e. the task is framed as an optimization problem rather than a data generating process.

Each type of identified migrant has a distinct configuration of individual, household and community characteristics.

Garip (2012) outlines the steps in cluster analysis as choosing:

1. the relevant attributes
2. an algorithm

After dividing the units, separate regressions are estimated for each cluster.
Garip (2012) uses the Mexican Migration Project data and discovers four types of migrants.

The approach is algorithmic rather than probabilistic, i.e. the task is framed as an optimization problem rather than a data generating process.

Each type of identified migrant has a distinct configuration of individual, household and community characteristics.

Garip (2012) outlines the steps in cluster analysis as choosing:

1. the relevant attributes
2. an algorithm
3. a similarity measure

After dividing the units, separate regressions are estimated for each cluster.
Garip (2012) uses the Mexican Migration Project data and discovers four types of migrants.

The approach is algorithmic rather than probabilistic, i.e. the task is framed as an optimization problem rather than a data generating process.

Each type of identified migrant has a distinct configuration of individual, household and community characteristics.

Garip (2012) outlines the steps in cluster analysis as choosing:

1. the relevant attributes
2. an algorithm
3. a similarity measure
4. number of clusters or mixture components
Garip (2012) uses the Mexican Migration Project data and discovers four types of migrants.

The approach is algorithmic rather than probabilistic, i.e. the task is framed as an optimization problem rather than a data generating process.

Each type of identified migrant has a distinct configuration of individual, household and community characteristics.

Garip (2012) outlines the steps in cluster analysis as choosing:

1. the relevant attributes
2. an algorithm
3. a similarity measure
4. number of clusters or mixture components
5. validation strategy
Garip (2012) uses the Mexican Migration Project data and discovers four types of migrants. The approach is algorithmic rather than probabilistic, i.e. the task is framed as an optimization problem rather than a data generating process. Each type of identified migrant has a distinct configuration of individual, household and community characteristics.

Garip (2012) outlines the steps in cluster analysis as choosing:
1. the relevant attributes
2. an algorithm
3. a similarity measure
4. number of clusters or mixture components
5. validation strategy

After dividing the units, separate regressions are estimated for each cluster.
Garip (2012) uses an iterative clustering algorithm related to $k$-means which minimizes the sum of distances between data points and a cluster center.
Garip (2012) uses an iterative clustering algorithm related to $k$-means which minimizes the sum of distances between data points and a cluster center.

What do algorithmic methods like $k$-means assume about the data?
Garip (2012) uses an iterative clustering algorithm related to $k$-means which minimizes the sum of distances between data points and a cluster center.

What do algorithmic methods like $k$-means assume about the data?

$k$-means assumes a distance metric and an objective function. This has a close connection to a probabilistic model. Different assumptions, but same underlying idea.
• Garip (2012) uses an iterative clustering algorithm related to \( k \)-means which minimizes the sum of distances between data points and a cluster center.

• What do algorithmic methods like \( k \)-means assume about the data?

• \( k \)-means assumes a distance metric and an objective function. This has a close connection to a probabilistic model. Different assumptions, but same underlying idea.

• Garip (2012) uses the “city block” or Manhattan distance which minimizes \( L_1 \) distance rather than the Euclidean distance.
Connections: $k$-means and Gaussian Mixtures

We started class with the example of a mixture model with Normally distributed components, often called a Gaussian Mixture Model (GMM). $k$-means typically minimizes the $L^2$ (Euclidean distance) which shares the squared-loss objective with the Gaussian distribution. We can obtain a correspondence between the two using small-variance asymptotics. As the covariances of the Gaussian go to zero, the EM algorithm for the GMM; $k$-means (Banerjee et al 2005, Kulis and Jordan 2012).

There is often a correspondence between probabilistic models and popular distance-based algorithms. This emphasizes the connections between assumptions about a distance or loss function and an assumption about the model.
Connections: \textit{k}-means and Gaussian Mixtures

- We started class with the example of a mixture model with Normally distributed components, often called a Gaussian Mixture Model (GMM)
Connections: \textit{k}-means and Gaussian Mixtures

- We started class with the example of a mixture model with Normally distributed components, often called a Gaussian Mixture Model (GMM).
- \textit{k}-means typically minimizes the $L_2$ (Euclidean distance) which shares the squared-loss objective with the Gaussian distribution.
Connections: $k$-means and Gaussian Mixtures

- We started class with the example of a mixture model with Normally distributed components, often called a Gaussian Mixture Model (GMM).
- $k$-means typically minimizes the $L_2$ (Euclidean distance) which shares the squared-loss objective with the Gaussian distribution.
- We can obtain a correspondence between the two using small-variance asymptotics. As the covariances of the Gaussian go to zero, the EM algorithm for the GMM $\leadsto k$-means (Banerjee et al 2005, Kulis and Jordan 2012).
Connections: \textit{k}-means and Gaussian Mixtures

- We started class with the example of a mixture model with Normally distributed components, often called a Gaussian Mixture Model (GMM)
- \textit{k}-means typically minimizes the $L_2$ (Euclidean distance) which shares the squared-loss objective with the Gaussian distribution.
- We can obtain a correspondence between the two using small-variance asymptotics. As the covariances of the Gaussian go to zero, the EM algorithm for the GMM $\rightsquigarrow$ \textit{k}-means (Banerjee et al 2005, Kulis and Jordan 2012).
- There is often a correspondence between probabilistic models and popular distance-based algorithms.
Connections: $k$-means and Gaussian Mixtures

- We started class with the example of a mixture model with Normally distributed components, often called a Gaussian Mixture Model (GMM).

- $k$-means typically minimizes the $L_2$ (Euclidean distance) which shares the squared-loss objective with the Gaussian distribution.

- We can obtain a correspondence between the two using small-variance asymptotics. As the covariances of the Gaussian go to zero, the EM algorithm for the GMM $\leadsto k$-means (Banerjee et al 2005, Kulis and Jordan 2012).

- There is often a correspondence between probabilistic models and popular distance-based algorithms.

- This emphasizes the connections between an assumptions about a distance or loss function and an assumption about the model.
Connections: \( k \)-means and Gaussian Mixtures

The biggest impact is that \( k \)-means strongly prefers equal sized clusters.
Connections: \textit{k}-means and Gaussian Mixtures

The biggest impact is that \textit{k}-means strongly prefers equal sized clusters.

\textbf{Different cluster analysis results on "mouse" data set:}

Original Data

\textbf{k-Means Clustering}

\textbf{EM Clustering}
Results

- Discovers four clusters and labels them: Income Maximizers, Risk Diversifiers, Network Migrants, Urban Migrants

- Estimates regressions for each of the four groups separately.

- Examines temporal trends for each (e.g. income maximizers come in early 1970s but decline over time).

- Finds that time trends in migrant types track closely with the introduction of new theory, i.e. theory describes the dominant empirical trend at the time of introduction.

- Big advance in our understanding with a data-driven approach!
Results

- Discovers four clusters and labels them: Income Maximizers, Risk Diversifiers, Network Migrants, Urban Migrants
- Estimates regressions for each of the four groups separately.
Results

- Discovers four clusters and labels them: Income Maximizers, Risk Diversifiers, Network Migrants, Urban Migrants
- Estimates regressions for each of the four groups separately.
- Examines temporal trends for each (e.g. income maximizers come in early 1970s but decline over time).
Results

- Discovers four clusters and labels them: Income Maximizers, Risk Diversifiers, Network Migrants, Urban Migrants.
- Estimates regressions for each of the four groups separately.
- Examines temporal trends for each (e.g. income maximizers come in early 1970s but decline over time).
- Finds that time trends in migrant types track closely with the introduction of new theory, i.e. theory describes the dominant empirical trend at the time of introduction.
Results

- Discovers four clusters and labels them: Income Maximizers, Risk Diversifiers, Network Migrants, Urban Migrants.
- Estimates regressions for each of the four groups separately.
- Examines temporal trends for each (e.g. income maximizers come in early 1970s but decline over time).
- Finds that time trends in migrant types track closely with the introduction of new theory, i.e. theory describes the dominant empirical trend at the time of introduction.
- Big advance in our understanding with a data-driven approach!
Next Steps

Garip (2012) uses clustering as a tool for discovery. Can the tool (clustering and regressing) be refined further? Is there a tool that incorporates uncertainty in the clustering and does not encourage equal-sized clusters explicitly clusters heterogeneity in migrant mechanisms instead of heterogeneity in migrant characteristics. Is there a model optimized for finding heterogeneous mechanisms?
Next Steps

Garip (2012) uses clustering as a tool for discovery.
Garip (2012) uses clustering as a tool for discovery.

Can the tool (clustering and regressing) be refined further? Is there a tool that...
Next Steps

Garip (2012) uses clustering as a tool for discovery.

Can the tool (clustering and regressing) be refined further?

Is there a tool that...

- incorporates uncertainty in the clustering
Next Steps

Garip (2012) uses clustering as a tool for discovery.

Can the tool (clustering and regressing) be refined further? Is there a tool that...

- incorporates uncertainty in the clustering
- does not encourage equal-sized clusters
Next Steps

Garip (2012) uses clustering as a tool for discovery. Can the tool (clustering and regressing) be refined further? Is there a tool that...

- incorporates uncertainty in the clustering
- does not encourage equal-sized clusters
- explicitly clusters heterogeneity in migrant mechanisms instead of heterogeneity in migrant characteristics.
Next Steps

Garip (2012) uses clustering as a tool for discovery.

Can the tool (clustering and regressing) be refined further?
Is there a tool that...
- incorporates uncertainty in the clustering
- does not encourage equal-sized clusters
- explicitly clusters heterogeneity in migrant mechanisms instead of heterogeneity in migrant characteristics.

Is there a model optimized for finding heterogeneous mechanisms?
Mixtures of Regressions

- Like Garip (2012), Imai and Tingley (2012) criticize the competing variable regression approach to theory testing.
Mixtures of Regressions

- Like Garip (2012), Imai and Tingley (2012) criticize the competing variable regression approach to theory testing.
- Imai and Tingley turn to mixtures of regressions, each observation is explained by one of $K$ regression models (or more generally all observations are defined by a common set of weights $\pi$).
Like Garip (2012), Imai and Tingley (2012) criticize the competing variable regression approach to theory testing.

Imai and Tingley turn to mixtures of regressions, each observation is explained by one of $K$ regression models (or more generally all observations are defined by a common set of weights $\pi$).

Each of these regressions can have the same or different sets of explanatory variables.
Like Garip (2012), Imai and Tingley (2012) criticize the competing variable regression approach to theory testing.

Imai and Tingley turn to mixtures of regressions, each observation is explained by one of $K$ regression models (or more generally all observations are defined by a common set of weights $\pi$).

Each of these regressions can have the same or different sets of explanatory variables.

Thus we have the log-likelihood

$$\ell = \sum_{i=1}^{N} \log \left( \sum_{k=1}^{K} \pi_k f_k(Y_i|X_i, \theta_k) \right)$$
The Applied Problem

- Hiscox (2002) wants to explain legislative voting on trade bills
The Applied Problem

- Hiscox (2002) wants to explain legislative voting on trade bills
- Two canonical theories:
The Applied Problem

- Hiscox (2002) wants to explain legislative voting on trade bills
- Two canonical theories:
  - Stolper-Samuelson (SS) \(\leadsto\) factor owners will support trade liberalization
  - Ricardo-Viner (RV)
The Applied Problem

- Hiscox (2002) wants to explain legislative voting on trade bills
- Two canonical theories:
  - Stolper-Samuelson (SS) \(\sim\) factor owners will support trade liberalization
  - Ricardo-Viner (RV) \(\sim\) exporting sectors will support trade liberalization
The Applied Problem

- Hiscox (2002) wants to explain legislative voting on trade bills
- Two canonical theories:
  - Stolper-Samuelson (SS) \(\rightsquigarrow\) factor owners will support trade liberalization
  - Ricardo-Viner (RV) \(\rightsquigarrow\) exporting sectors will support trade liberalization
- Hiscox (2002) hypothesis that voting explained on the factor specificity in the U.S. economy at that time
The Applied Problem

- Hiscox (2002) wants to explain legislative voting on trade bills
- Two canonical theories:
  - Stolper-Samuelson (SS) \( \sim \) factor owners will support trade liberalization
  - Ricardo-Viner (RV) \( \sim \) exporting sectors will support trade liberalization
- Hiscox (2002) hypothesis that voting explained on the factor specificity in the U.S. economy at that time
- Test by dividing up data in time and shows that liberalization best accounted for by SS when specificity is low, reverse for RV
The Applied Problem

- Hiscox (2002) wants to explain legislative voting on trade bills
- Two canonical theories:
  - Stolper-Samuelson (SS) $\rightsquigarrow$ factor owners will support trade liberalization
  - Ricardo-Viner (RV) $\rightsquigarrow$ exporting sectors will support trade liberalization
- Hiscox (2002) hypothesis that voting explained on the factor specificity in the U.S. economy at that time
- Test by dividing up data in time and shows that liberalization best accounted for by SS when specificity is low, reverse for RV
- Any one division in time open to critique- can we do better?
Testing Competing Theories

- We can now specify a mixture of two regressions, one for each theory (RV and SS)
Testing Competing Theories

- We can now specify a mixture of two regressions, one for each theory (RV and SS)
- To test Hiscox theory that the choice is driven by factor specificity, we can parameterize the indicator $Z_i$ as:

$$P(Z_i = m|W_i) = \pi_m(W_i, \psi_m)$$

After fitting the model we know how well each theory predicts each observation, as well as what covariates are associated with that theory. They find that evidence for Hiscox's hypothesis is fairly weak and more data is necessary for a strong test. They also find more interpretable results with all coefficients in the expected directions from the theory.
Testing Competing Theories

- We can now specify a mixture of two regressions, one for each theory (RV and SS)
- To test Hiscox theory that the choice is driven by factor specificity, we can parameterize the indicator $Z_i$ as:

$$ P(Z_i = m|W_i) = \pi_m(W_i, \psi_m) $$

- After fitting the model we know how well each theory predicts each observation, as well as what covariates are associated with that theory choice.
Testing Competing Theories

- We can now specify a mixture of two regressions, one for each theory (RV and SS).
- To test Hiscox theory that the choice is driven by factor specificity, we can parameterize the indicator $Z_i$ as:

$$P(Z_i = m|W_i) = \pi_m(W_i, \psi_m)$$

- After fitting the model we know how well each theory predicts each observation, as well as what covariates are associated with that theory choice.
- They find that evidence for Hiscox’s hypothesis is fairly weak and more data is necessary for a strong test.
Testing Competing Theories

- We can now specify a mixture of two regressions, one for each theory (RV and SS)
- To test Hiscox theory that the choice is driven by factor specificity, we can parameterize the indicator $Z_i$ as:

$$P(Z_i = m | W_i) = \pi_m(W_i, \psi_m)$$

- After fitting the model we know how well each theory predicts each observation, as well as what covariates are associated with that theory choice.
- They find that evidence for Hiscox’s hypothesis is fairly weak and more data is necessary for a strong test.
- They also find more interpretable results with all coefficients in the expected directions from the theory.
Pitfalls in Mixture Modeling

- **Discovered** groups don’t necessarily correspond to a **desired** latent indicator (e.g. the height example)
Pitfalls in Mixture Modeling

- **Discovered** groups don’t necessarily correspond to a *desired* latent indicator (e.g. the height example)
- The models are *not identified* due to label-switching
Pitfalls in Mixture Modeling

- Discovered groups don’t necessarily correspond to a desired latent indicator (e.g. the height example)
- The models are not identified due to label-switching
- Even beyond label-switching, the likelihoods have multiple local maxima
Pitfalls in Mixture Modeling

- **Discovered** groups don’t necessarily correspond to a **desired** latent indicator (e.g. the height example)
- The models are **not identified** due to label-switching
- Even beyond label-switching, the likelihoods have multiple **local maxima**
- **Estimation** is difficult and the likelihood can have infinite spikes
Pitfalls in Mixture Modeling

- **Discovered** groups don’t necessarily correspond to a **desired** latent indicator (e.g. the height example)
- The models are **not identified** due to label-switching
- Even beyond label-switching, the likelihoods have multiple **local maxima**
- **Estimation** is difficult and the likelihood can have infinite spikes
- It is difficult to choose the number of **clusters/components**
The Promise of Mixture Modeling

- We can discover latent groups giving us new theoretical insights, methods to test theories, and discovery of heterogeneity
The Promise of Mixture Modeling

- We can discover latent groups giving us new theoretical insights, methods to test theories, and discovery of heterogeneity
- Mixtures are more flexible models of complex distributions
The Promise of Mixture Modeling

- We can discover latent groups giving us new theoretical insights, methods to test theories, and discovery of heterogeneity.
- Mixtures are more flexible models of complex distributions.
- The mixture infrastructure is modular and can be plugged into many other model setups.
1. Mixture Models
   - Basic Mixtures
   - Application: Mixtures as Preprocessing
   - Application: Mixture of Regressions

2. Expectation Maximization
   - EM for Probit Regression
   - EM for Gaussian Mixtures
   - EM in General

3. Missing Data
   - Motivating Example
   - Overview and Assumptions
   - Existing Heuristics
   - Application Specific Approaches
   - Multiple Imputation
   - The Full Amelia Scheme

4. Measurement Error

5. Appendix: Additional Details and Examples
Overview

Expectation-Maximization (EM) is a very general algorithm for maximizing a likelihood in the presence of missing data. Often we use it when the missingness comes from data augmentation where we introduce a latent variable to make computation more straightforward.

Core Idea:

▶ if we knew the latent variable estimating the model parameters would be easy,
▶ if we knew the model parameters estimating the latent variables would be easy.

EM has two steps which are iterated:

▶ E-Step: update the latent variables by taking the expectation
▶ M-Step: update the model parameters by maximizing the complete data likelihood

We will step through a few cases to see how this works.
Overview

- Expectation-Maximization (EM) is a very general algorithm for maximizing a likelihood in the presence of missing data.

Core Idea:
- if we knew the latent variable estimating the model parameters would be easy,
- if we knew the model parameters estimating the latent variables would be easy.

EM has two steps which are iterated:
- E-Step: update the latent variables by taking the expectation
- M-Step: update the model parameters by maximizing the complete data likelihood

We will step through a few cases to see how this works.
Overview

- Expectation-Maximization (EM) is a very general algorithm for maximizing a likelihood in the presence of missing data.
- Often we use it when the missingness comes from data augmentation where we introduce a latent variable to make computation more straightforward.
Overview

- Expectation-Maximization (EM) is a very general algorithm for maximizing a likelihood in the presence of missing data.
- Often we use it when the missingness comes from data augmentation where we introduce a latent variable to make computation more straightforward.
- Core Idea:
  - if we knew the latent variable estimating the model parameters would be easy,
Overview

- Expectation-Maximization (EM) is a very general algorithm for maximizing a likelihood in the presence of missing data.
- Often we use it when the missingness comes from data augmentation where we introduce a latent variable to make computation more straightforward.
- Core Idea:
  - if we knew the latent variable estimating the model parameters would be easy,
  - if we knew the model parameters estimating the latent variables would be easy.
Overview

- Expectation-Maximization (EM) is a very general algorithm for maximizing a likelihood in the presence of missing data.

- Often we use it when the missingness comes from data augmentation where we introduce a latent variable to make computation more straightforward.

- Core Idea:
  - if we knew the latent variable estimating the model parameters would be easy,
  - if we knew the model parameters estimating the latent variables would be easy.

- EM has two steps which are iterated:
  - E-Step: update the latent variables by taking the expectation
  - M-Step: update the model parameters by maximizing the complete data likelihood
Overview

- Expectation-Maximization (EM) is a very general algorithm for maximizing a likelihood in the presence of missing data.
- Often we use it when the missingness comes from data augmentation where we introduce a latent variable to make computation more straightforward.

**Core Idea:**

- if we knew the latent variable estimating the model parameters would be easy,
- if we knew the model parameters estimating the latent variables would be easy.

**EM has two steps which are iterated:**

- **E-Step:** update the latent variables by taking the expectation.
Expectation-Maximization (EM) is a very general algorithm for maximizing a likelihood in the presence of missing data. Often we use it when the missingness comes from data augmentation where we introduce a latent variable to make computation more straightforward.

Core Idea:
- If we knew the latent variable estimating the model parameters would be easy,
- If we knew the model parameters estimating the latent variables would be easy.

EM has two steps which are iterated:
- **E-Step**: update the latent variables by taking the expectation
- **M-Step**: update the model parameters by maximizing the complete data likelihood
Overview

- Expectation-Maximization (EM) is a very general algorithm for maximizing a likelihood in the presence of missing data.
- Often we use it when the missingness comes from data augmentation where we introduce a latent variable to make computation more straightforward.
- Core Idea:
  - if we knew the latent variable estimating the model parameters would be easy,
  - if we knew the model parameters estimating the latent variables would be easy.
- EM has two steps which are iterated:
  - E-Step: update the latent variables by taking the expectation
  - M-Step: update the model parameters by maximizing the complete data likelihood
- We will step through a few cases to see how this works.
Let $Y_i^* \sim P(y_i^*|\mu_i)$ where $\mu_i = X_i\beta$ and assume that we only observe

$$Y_i = \begin{cases} 
1 & \text{if } y_i^* \geq \tau \\
0 & \text{if } y_i^* < \tau 
\end{cases}$$
Review of the Probit Latent Regression Formulation

Let $Y_i^* \sim P(y_i^*|\mu_i)$ where $\mu_i = X_i\beta$ and assume that we only observe

$$Y_i = \begin{cases} 
1 & \text{if } y_i^* \geq \tau \\
0 & \text{if } y_i^* < \tau
\end{cases}$$

For the probit model, $P(\cdot) = \mathcal{N}(\mu_i, \sigma^2)$. Typically assume that $\tau = 0$ and $\sigma = 1$ in order to fit the model.
What if we observed $Y_i^*$?

$$Y_i^* = X\beta + \epsilon_i$$
The Intuition

What if we observed $Y_i^*$?

$$Y_i^* = X\beta + \epsilon_i$$

How do we estimate $\beta$?
The Intuition

What if we observed $Y_i^*$?

$$Y_i^* = X\beta + \epsilon_i$$

How do we estimate $\beta$?

$$\hat{\beta} = (X'X)^{-1}X'Y^*$$
The Intuition

What if we observed $Y_i^*$?

$$Y_i^* = X \beta + \epsilon_i$$

How do we estimate $\beta$?

$$\hat{\beta} = (X'X)^{-1}X'Y^*$$

But oh yeah, we don’t know $Y_i^*$
The Intuition

What if we knew $\beta$?

$$Y_i^* = X_i \beta + \epsilon_i$$

We'll come back to that last part in a second.
The Intuition

What if we knew $\beta$?

$$Y_i^* = X_i \beta + \epsilon_i$$

We still wouldn’t know $Y_i^*$ but we could calculate $E(Y_i^* | y_i, X_i, \beta)$.
The Intuition

What if we knew \( \beta \)?

\[ Y_i^* = X_i \beta + \epsilon_i \]

We still wouldn’t know \( Y_i^* \) but we could calculate \( E(Y_i^*|y_i, X_i, \beta) \)

\[
E(Y_i^*|y_i, X_i, \beta) = E(X_i \beta + \epsilon_i|y_i, X_i, \beta)
\]

\[
= E(X_i \beta|y_i, X_i, \beta) + E(\epsilon_i|y_i, X_i, \beta)
\]

\[
= X_i \beta + E(\epsilon_i|y_i, X_i, \beta)
\]

We’ll come back to that last part in a second.
The Intuition

What if we knew $\beta$?

$$Y_i^* = X_i \beta + \epsilon_i$$

We still wouldn’t know $Y_i^*$ but we could calculate $E(Y_i^*|y_i, X_i, \beta)$

$$E(Y_i^*|y_i, X_i, \beta) = E(X_i \beta + \epsilon_i|y_i, X_i, \beta)$$

$$= E(X_i \beta|y_i, X_i, \beta) + E(\epsilon_i|y_i, X_i, \beta)$$

$$= X_i \beta + E(\epsilon_i|y_i, X_i, \beta)$$

We’ll come back to that last part in a second.
The Intuition

This suggests an iterative procedure where we make up some data (called data augmentation). So we start with some plausible initial values of $\beta$ which we will call $\beta^t$. 

1. **E-Step**
   Take the expectation of the latent variable conditional on the current value of the parameters to impute the missing data.

   $y^*_t$,
   $y^*_i = E(Y^*_i | y_i, X_i, \beta_t)$

2. **M-Step**
   Maximize the complete data log-likelihood.

   $\beta^{(t+1)} = (X'X)^{-1}X'y^*_t$. 

3. Increment until convergence.
The Intuition

This suggests an iterative procedure where we make up some data (called data augmentation). So we start with some plausible initial values of $\beta$ which we will call $\beta^t$.

1. **E-Step** Take the expectation of the latent variable conditional on the current value of the parameters to impute the missing data.

$$y_{i}^{*,t} = E(Y_i^*|y_i, X_i, \beta^t)$$
This suggests an iterative procedure where we make up some data (called data augmentation). So we start with some plausible initial values of $\beta$ which we will call $\beta^t$.

1. **E-Step** Take the expectation of the latent variable conditional on the current value of the parameters to impute the missing data.
   \[ y_{i^*,t} = E(Y_{i^*}|y_i, X_i, \beta^t) \]

2. **M-Step** Maximize the complete data log-likelihood.
   \[ \beta^{(t+1)} = (X'X)^{-1}X'y_{*,t} \]
The Intuition

This suggests an iterative procedure where we make up some data (called data augmentation). So we start with some plausible initial values of $\beta$ which we will call $\beta^t$.

1. **E-Step** Take the expectation of the latent variable conditional on the current value of the parameters to impute the missing data.
   
   $y_{i,*}^{*,t} = E( Y_{i,*} | y_i, X_i, \beta^t )$

2. **M-Step** Maximize the complete data log-likelihood.
   
   $\beta^{(t+1)} = (X'X)^{-1} X' y_{*,*}^{*,t}$.

3. Increment until convergence.
The EM Algorithm

This is called the EM (Expectation-Maximization) Algorithm. It is due to Dempster, Laird and Rubin 1977.

Some Useful Facts:
1. This is a mode finding algorithm so it will retrieve the exact maximum likelihood estimates.
2. Each step will generate a higher (or constant) likelihood.
3. It is guaranteed to converge under very general conditions.
The EM Algorithm

This is called the EM (Expectation-Maximization) Algorithm. It is due to Dempster, Laird and Rubin 1977.
The EM Algorithm

This is called the EM (Expectation-Maximization) Algorithm. It is due to Dempster, Laird and Rubin 1977.

Some Useful Facts:

1. This is a mode finding algorithm so it will retrieve the exact maximum likelihood estimates.
2. Each step will generate a higher (or constant) likelihood.
3. It is guaranteed to converge under very general conditions.
The EM Algorithm

This is called the EM (Expectation-Maximization) Algorithm. It is due to Dempster, Laird and Rubin 1977.

Some Useful Facts:

1. This is a mode finding algorithm so it will retrieve the exact maximum likelihood estimates.
The EM Algorithm

This is called the EM (Expectation-Maximization) Algorithm. It is due to Dempster, Laird and Rubin 1977.

Some Useful Facts:

1. This is a mode finding algorithm so it will retrieve the exact maximum likelihood estimates.
2. Each step will generate a higher (or constant) likelihood.
The EM Algorithm

This is called the EM (Expectation-Maximization) Algorithm. It is due to Dempster, Laird and Rubin 1977.

Some Useful Facts:

1. This is a mode finding algorithm so it will retrieve the exact maximum likelihood estimates.
2. Each step will generate a higher (or constant) likelihood.
3. It is guaranteed to converge under very general conditions.
The EM Algorithm for the Probit Case

So we know our algorithm has two major steps. But what are they?

1. Posit some initial values of $\beta_t$
2. Calculate the $E(Y^*_{i|y_i, X_i, \beta_t})$
   
   $E(Y^*_{i|y_i, X_i, \beta_t}) = E(X_i \beta_t + \epsilon_i|y_i, X_i, \beta_t)$
   
   $= X_i \beta_t + E(\epsilon_i|y_i, X_i, \beta_t)$
   
   $= X_i \beta_t + ( - \phi_i(-X_i \beta_t)(1 - y_i) \Phi_i(-X_i \beta_t)$

3. Calculate the estimate for $\beta_{t+1}$ using the complete data.
   
   $\hat{\beta}_{t+1} = (X'X)^{-1}X'E(Y^*_{i|y_i, X_i, \beta_t})$

4. Repeat Steps 2-3 Until Convergence.

Note that the $E(\epsilon_i)$ is related to the truncated normal, because we have information about the sign from $y_i$. 
The EM Algorithm for the Probit Case

So we know our algorithm has two major steps. But what are they?

1. Posit some initial values of $\beta^t$

2. Calculate the $E(Y^*_i | y_i, X_i, \beta^t)$

$$E(Y^*_i | y_i, X_i, \beta^t) = E(X_i \beta^t + \epsilon_i | y_i, X_i, \beta^t) = X_i \beta^t + E(\epsilon_i | y_i, X_i, \beta^t)$$

$$= X_i \beta^t + (-\phi_i (-X_i \beta^t)) \Phi_i (-X_i \beta^t) (1-y_i)((1-\Phi_i (-X_i \beta^t)))$$

3. Calculate the estimate for $\beta^{t+1}$ using the complete data.

$$\hat{\beta}^{t+1} = (X^t X)^{-1} X^t E(Y^*_i | y_i, X_i, \beta^t)$$

4. Repeat Steps 2-3 Until Convergence.

Note that the $E(\epsilon_i)$ is related to the truncated normal, because we have information about the sign from $y_i$. 

Stewart (Princeton)  
Missing Data  
Mar 27-Apr 5, 2017 34 / 200
The EM Algorithm for the Probit Case

So we know our algorithm has two major steps. But what are they?

1. Posit some initial values of $\beta^t$
2. Calculate the $E(Y^*_i|y_i, X_i, \beta^t)$

$$E(Y^*_i|y_i, X_i, \beta^t) = E(X_i\beta^t + \epsilon_i|y_i, X_i, \beta)$$

$$= X_i\beta^t + E(\epsilon_i|y_i, X_i, \beta)$$

$$= X_i\beta^t + \left( \frac{-\phi_i(-X_i\beta^t)}{\Phi_i(-X_i\beta^t)} \right)^{(1-y_i)} \left( \frac{\phi_i(-X_i\beta^t)}{(1 - \Phi_i(-X_i\beta^t))} \right)^{y_i}$$

Note that the $E(\epsilon_i)$ is related to the truncated normal, because we have information about the sign from $y_i$. 
The EM Algorithm for the Probit Case

So we know our algorithm has two major steps. But what are they?

1. Posit some initial values of $\beta^t$
2. Calculate the $E(Y_i^*|y_i, X_i, \beta^t)$

$$E(Y_i^*|y_i, X_i, \beta^t) = E(X_i\beta^t + \epsilon_i|y_i, X_i, \beta)$$
$$= X_i\beta^t + E(\epsilon_i|y_i, X_i, \beta)$$

$$= X_i\beta^t + \left(\frac{-\phi_i(-X_i\beta^t)}{\Phi_i(-X_i\beta^t)}\right)^{(1-y_i)} \left(\frac{\phi_i(-X_i\beta^t)}{(1 - \Phi_i(-X_i\beta^t))}\right)^{y_i}$$

3. Calculate the estimate for $\beta^{t+1}$ using the complete data.

$$\hat{\beta}^{(t+1)} = (X'X)^{-1}X'E(Y_i^*|y_i, X_i, \beta^t)$$

Note that the $E(\epsilon_i)$ is related to the truncated normal, because we have information about the sign from $y_i$. 
The EM Algorithm for the Probit Case

So we know our algorithm has two major steps. But what are they?

1. Posit some initial values of $\beta^t$

2. Calculate the $E(Y_i^*|y_i, X_i, \beta^t)$

$$E(Y_i^*|y_i, X_i, \beta^t) = E(X_i\beta^t + \epsilon_i|y_i, X_i, \beta)$$

$$= X_i\beta^t + E(\epsilon_i|y_i, X_i, \beta)$$

$$= X_i\beta^t + \left(\frac{-\phi_i(-X_i\beta^t)}{\Phi_i(-X_i\beta^t)}\right)^{(1-y_i)} \left(\frac{\phi_i(-X_i\beta^t)}{(1 - \Phi_i(-X_i\beta^t))}\right)^{y_i}$$

3. Calculate the estimate for $\beta^{t+1}$ using the complete data.

$$\hat{\beta}^{(t+1)} = (X'X)^{-1}X'E(Y_i^*|y_i, X_i, \beta^t)$$

4. Repeat Steps 2-3 Until Convergence.

Note that the $E(\epsilon_i)$ is related to the truncated normal, because we have information about the sign from $y_i$. 
The EM Algorithm

Note that this is a really high-level, heuristic view of EM. The steps are always the same though:
The EM Algorithm

Note that this is a really high-level, heuristic view of EM. The steps are always the same though:

1. Identify the latent variables $Z$ and the parameters $\theta$. 
The EM Algorithm

Note that this is a really high-level, heuristic view of EM. The steps are always the same though:

1. Identify the latent variables $Z$ and the parameters $\theta$.
2. Identify the target density (called the Q function)
The EM Algorithm

Note that this is a really high-level, heuristic view of EM. The steps are always the same though:

1. Identify the latent variables $Z$ and the parameters $\theta$.
2. Identify the target density (called the Q function)

$$Q(\theta, \theta^{(t)}) = \int p(Z|\theta^{(t)}, Y) \log p(Z, Y|\theta) dZ$$
The EM Algorithm

Note that this is a really high-level, heuristic view of EM. The steps are always the same though:

1. Identify the latent variables $Z$ and the parameters $\theta$.
2. Identify the target density (called the Q function)

$$Q(\theta, \theta^{(t)}) = \int p(Z|\theta^{(t)}, Y)\log p(Z, Y|\theta) dZ$$

3. E-step: compute $Z^{(t)} = E(Z|\theta^{(t)}, Y)$


$$\theta^{(t+1)} = \text{argmax}_{\theta} Q(\theta, \theta^{(t)})$$

5. Assess convergence either by changes in parameters or the log-likelihood.
The EM Algorithm

Note that this is a really high-level, heuristic view of EM. The steps are always the same though:

1. Identify the latent variables \( Z \) and the parameters \( \theta \).
2. Identify the target density (called the Q function)

\[
Q(\theta, \theta^{(t)}) = \int p(Z|\theta^{(t)}, Y) \log p(Z, Y|\theta) dZ
\]

3. E-step: compute \( Z^{(t)} = E(Z|\theta^{(t)}, Y) \)
\[
\theta^{(t+1)} = \arg \max_{\theta} Q(\theta, \theta^{t})
\]
The EM Algorithm

Note that this is a really high-level, heuristic view of EM. The steps are always the same though:

1. Identify the latent variables $Z$ and the parameters $\theta$.
2. Identify the target density (called the Q function)
   \[ Q(\theta, \theta^{(t)}) = \int p(Z|\theta^{(t)}, Y) \log p(Z, Y|\theta) dZ \]
3. E-step: compute $Z^{(t)} = E(Z|\theta^{(t)}, Y)$
4. M-step: maximize the complete data log-likelihood. $\theta^{(t+1)} = \arg\max_\theta Q(\theta, \theta^{t})$
5. Assess convergence either by changes in parameters or the log-likelihood.
Example 2: Mixtures

Single distribution data generating process:
Example 2: Mixtures

Single distribution data generating process:

\[ x_i \sim \text{Distribution}(\text{parameters}) \]
Example 2: Mixtures

Single distribution data generating process:

\[ x_i \sim \text{Distribution(parameters)} \]

Mixture of distribution data generating process:

[\begin{align*}
  z_i & \sim \text{Multinomial}(1, \pi) \\
  x_i | z_i = k & \sim \text{Distribution}(\text{parameters}_k)
\end{align*}]
Example 2: Mixtures

Single distribution data generating process:

\[ x_i \sim \text{Distribution}(\text{parameters}) \]

Mixture of distribution data generating process:

\[ z_i | \pi \sim \text{Multinomial}(1, \pi) \]

In words:

Draw a cluster label

Given distribution, draw realization
Example 2: Mixtures

Single distribution data generating process:

\[ x_i \sim \text{Distribution}(\text{parameters}) \]

Mixture of distribution data generating process:

\[ z_i | \pi \sim \text{Multinomial}(1, \pi) \]
\[ x_i | z_{ik} = 1 \sim \text{Distribution}(\text{parameters}_k) \]
Example 2: Mixtures

Single distribution data generating process:

\[ x_i \sim \text{Distribution(parameters)} \]

Mixture of distribution data generating process:

\[ z_i \mid \pi \sim \text{Multinomial}(1, \pi) \]
\[ x_i \mid z_{ik} = 1 \sim \text{Distribution(parameters}_k) \]

In words:

Draw a cluster label
Given distribution, draw realization
Example 2: Mixtures

Single distribution data generating process:

\[ x_i \sim \text{Distribution(parameters)} \]

Mixture of distribution data generating process:

\[ z_i \mid \pi \sim \text{Multinomial}(1, \pi) \]
\[ x_i \mid z_{ik} = 1 \sim \text{Distribution(parameters}_k) \]

In words:
- Draw a cluster label
Example 2: Mixtures

Single distribution data generating process:

\[ x_i \sim \text{Distribution}(\text{parameters}) \]

Mixture of distribution data generating process:

\[ z_i | \pi \sim \text{Multinomial}(1, \pi) \]
\[ x_i | z_{ik} = 1 \sim \text{Distribution}(\text{parameters}_k) \]

In words:

- Draw a cluster label
- Given distribution, draw realization
Gaussian Mixture

This leads to the likelihood:

\[ p(x) = \sum_z p(z) p(x | z) \]

\[ = K \sum_{k=1}^{\text{K}} \pi_k N(x | \mu_k, \Sigma_k) \]
Gaussian Mixture

\[ z_i | \pi \sim \text{Multinomial}(1, \pi) \]
Gaussian Mixture

\[ z_i | \pi \sim \text{Multinomial}(1, \pi) \]
\[ x_i | z_{ik} = 1, \mu_k, \Sigma_k \sim \text{Normal}(\mu_k, \Sigma_k) \]
Gaussian Mixture

\[ z_i | \pi \sim \text{Multinomial}(1, \pi) \]
\[ x_i | z_{ik} = 1, \mu_k, \Sigma_k \sim \text{Normal}(\mu_k, \Sigma_k) \]

This leads to the likelihood:

\[
p(x) = \sum_z p(z)p(x | z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x | \mu_k, \Sigma_k)
\]
Algorithm for the Gaussian Mixture

1) Initialize parameters $\mu^t, \Sigma^t, \pi^t$
Algorithm for the Gaussian Mixture

1) Initialize parameters $\mu^t, \Sigma^t, \pi^t$
2) Expectation step: compute ‘responsibilities’ $p(z_i|\mu^t, \Sigma^t, \pi^t, X) \sim r_i^t$
Algorithm for the Gaussian Mixture

1) Initialize parameters $\mu^t, \Sigma^t, \pi^t$

2) Expectation step: compute ‘responsibilities’ $p(z_i|\mu^t, \Sigma^t, \pi^t, X) \sim r_i^t$

$$r_{ik} = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(x_i|\mu_{k'}, \Sigma_{k'})}$$
Algorithm for the Gaussian Mixture

1) Initialize parameters $\mu^t, \Sigma^t, \pi^t$

2) Expectation step: compute ‘responsibilities’ $p(z_i | \mu^t, \Sigma^t, \pi^t, X) \sim r_i^t$

$$r_{ik} = \frac{\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(x_i | \mu_{k'}, \Sigma_{k'})}$$

3) Maximization step: maximize with respect to $\mu, \Sigma$ and $\pi$: 
Algorithm for the Gaussian Mixture

1) Initialize parameters $\mu^t, \Sigma^t, \pi^t$

2) Expectation step: compute ‘responsibilities’ $p(z_i | \mu^t, \Sigma^t, \pi^t, X) \sim r^t_i$

$$r_{ik} = \frac{\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(x_i | \mu_{k'}, \Sigma_{k'})}$$

3) Maximization step: maximize with respect to $\mu, \Sigma$ and $\pi$:

$$E_z[\log p(x, z | \mu_k, \Sigma_k, \pi)] = E_z \left[ \log \left( \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(x_n | \mu_k, \Sigma_k)^{z_{nk}} \right) \right]$$

$$= E_z \left[ \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left[ \log \pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k) \right] \right]$$

Obtain $\mu_k^{t+1}, \Sigma_k^{t+1}, \pi^{t+1}$
Algorithm for the Gaussian Mixture

1) Initialize parameters $\mu^t, \Sigma^t, \pi^t$

2) **Expectation step**: compute ‘responsibilities’ $p(z_i|\mu^t, \Sigma^t, \pi^t, X) \sim r^t_i$

   $$r_{ik} = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(x_i|\mu_{k'}, \Sigma_{k'})}$$

3) **Maximization step**: maximize with respect to $\mu, \Sigma$ and $\pi$:

   $$E_z[\log p(x, z|\mu_k, \Sigma_k, \pi)] = E_z \left[ \log \left( \prod_{i=1}^N \prod_{k=1}^K \pi_k^{z_{nk}} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{nk}} \right) \right]$$

   $$= E_z \left[ \sum_{n=1}^N \sum_{k=1}^K z_{nk} \log \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k) \right]$$

   Obtain $\mu^{t+1}_k, \Sigma^{t+1}_k, \pi^{t+1}$

4) Assess change in the log-likelihood
Algorithm for the Gaussian Mixture

3) M-Step:

\[
\text{E}[\log \text{Complete data} | \theta, \pi] = \sum_{i=1}^{N} \sum_{k=1}^{K} \text{E}[z_{ik}] \log (\pi_k N(x_i | \mu_k, \Sigma_k))
\]

Because \( E[z_{ik}] = r_{ik} \), solutions are weighted averages of usual updates:

\[
\pi_{t+1}^k = \sum_{i=1}^{N} r_{tik} / N(1)
\]

\[
\mu_{t+1}^k = \frac{\sum_{i=1}^{N} r_{tik} x_i \sum_{i=1}^{N} r_{ik}}{\sum_{i=1}^{N} r_{ik}^2}
\]

\[
\Sigma_{t+1}^k = \frac{1}{\sum_{i=1}^{N} r_{ik}} \sum_{i=1}^{N} r_{ik} (x_i - \mu_{t+1}^k)(x_i - \mu_{t+1}^k)^T
\]
Algorithm for the Gaussian Mixture

3) M-Step:

\[
E[\log \text{Complete data}|\theta, \pi] = \sum_{i=1}^{N} \sum_{k=1}^{K} E[z_{ik}] \log (\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k))
\]
Algorithm for the Gaussian Mixture

3) M-Step:

$$E[\log \text{Complete data}|\theta, \pi] = \sum_{i=1}^{N} \sum_{k=1}^{K} E[z_{ik}] \log (\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k))$$

Because $E[z_{ik}] = r_{ik}$, solutions are weighted averages of usual updates.
Algorithm for the Gaussian Mixture

3) M-Step:

$$E[\log \text{Complete data}|\theta, \pi] = \sum_{i=1}^{N} \sum_{k=1}^{K} E[z_{ik}] \log (\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k))$$

Because $E[z_{ik}] = r_{ik}$, solutions are weighted averages of usual updates

$$\pi_{k}^{t+1} = \frac{\sum_{i=1}^{N} r_{ik}^t}{N}$$  \hspace{1cm} (1)$$
Algorithm for the Gaussian Mixture

3) M-Step:

$$E[\log \text{Complete data}|\theta, \pi] = \sum_{i=1}^{N} \sum_{k=1}^{K} E[z_{ik}] \log (\pi_k N(x_i|\mu_k, \Sigma_k))$$

Because $E[z_{ik}] = r_{ik}$, solutions are weighted averages of usual updates

$$\pi_{k}^{t+1} = \frac{\sum_{i=1}^{N} r_{ik}^{t}}{N}$$

$$\mu_{k}^{t+1} = \frac{\sum_{i=1}^{N} r_{ik}^{t} x_i}{\sum_{i=1}^{N} r_{ik}^{t}}$$
Algorithm for the Gaussian Mixture

3) M-Step:

$$E[\log \text{Complete data}|\theta, \pi] = \sum_{i=1}^{N} \sum_{k=1}^{K} E[z_{ik}] \log (\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k))$$

Because $E[z_{ik}] = r_{ik}$, solutions are weighted averages of usual updates

$$\pi_k^{t+1} = \frac{\sum_{i=1}^{N} r_{ik}^{t}}{N} \quad (1)$$

$$\mu_k^{t+1} = \frac{\sum_{i=1}^{N} r_{ik}^{t} x_i}{\sum_{i=1}^{N} r_{ik}^{t}} \quad (2)$$

$$\Sigma_k^{t+1} = \frac{1}{\sum_{i=1}^{N} r_{ik}^{t}} \sum_{i=1}^{N} r_{ik} (x_i - \mu_k^{t+1})(x_i - \mu_k^{t+1})^T \quad (3)$$
The EM Algorithm in Words

Consider a model for observed data $x$ that is accompanied by a latent $z$. A model with parameters $\theta$ describes the joint distribution of $x$ and $z$, as $p(x, z | \theta)$.

Under the maximum likelihood framework we want to find $\theta$ which maximizes:

$$p(x | \theta) = \int p(x, z | \theta) \, dz$$

We assume that maximizing the likelihood isn't easy but we can find $\theta$ to maximize $p(x, z | \theta)$ for known $x, z$. We know $x$ and so we plug in our best guess of $z$, the expectation.
Consider a model for observed data $x$ that is accompanied by a latent $z$. A model with parameters $\theta$ describes the joint distribution of $x$ and $z$, as $p(x, z|\theta)$. 
The EM Algorithm in Words

Consider a model for observed data $x$ that is accompanied by a latent $z$. A model with parameters $\theta$ describes the joint distribution of $x$ and $z$, as $p(x, z|\theta)$.

Under the maximum likelihood framework we want to find $\theta$ which maximizes:

$$p(x|\theta) = \int p(x, z|\theta)dz$$
The EM Algorithm in Words

Consider a model for observed data $x$ that is accompanied by a latent $z$. A model with parameters $\theta$ describes the joint distribution of $x$ and $z$, as $p(x, z|\theta)$.

Under the maximum likelihood framework we want to find $\theta$ which maximizes:

$$p(x|\theta) = \int p(x, z|\theta)dz$$

We assume that maximizing the likelihood isn’t easy but we can find $\theta$ to maximize $p(x, z|\theta)$ for known $x, z$. 

The EM Algorithm in Words

Consider a model for observed data $x$ that is accompanied by a latent $z$. A model with parameters $\theta$ describes the joint distribution of $x$ and $z$, as $p(x, z|\theta)$.

Under the maximum likelihood framework we want to find $\theta$ which maximizes:

$$p(x|\theta) = \int p(x, z|\theta)dz$$

We assume that maximizing the likelihood isn’t easy but we can find $\theta$ to maximize $p(x, z|\theta)$ for known $x, z$.

We know $x$ and so we plug in our best guess of $z$, the expectation.
The EM Algorithm in Math

1) Initialize parameters $\theta^t$
The EM Algorithm in Math

1) Initialize parameters $\theta^t$

2) E step: Using the current value of the parameter $\theta$ compute the expected value of the log-likelihood with respect to the conditional distribution of $Z|X$

$$Q(\theta|\theta^t) = E_{Z|X,\theta^t} [\log p(X, Z|\theta)]$$  (4)
The EM Algorithm in Math

1) Initialize parameters $\theta^t$

2) E step: Using the current value of the parameter $\theta$ compute the expected value of the log-likelihood with respect to the conditional distribution of $Z|X$

$$Q(\theta|\theta^t) = E_{Z|X,\theta^t} [\log p(X, Z|\theta)]$$

3) M step: maximize the Q function:
The EM Algorithm in Math

1) Initialize parameters $\theta^t$

2) E step: Using the current value of the parameter $\theta$ compute the expected value of the log-likelihood with respect to the conditional distribution of $Z|X$

$$Q(\theta|\theta^t) = E_{Z|X,\theta^t} [\log p(X, Z|\theta)]$$ (4)

3) M step: maximize the Q function:

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^t)$$
1) Initialize parameters $\theta^t$

2) E step: Using the current value of the parameter $\theta$ compute the expected value of the log-likelihood with respect to the conditional distribution of $Z|X$

$$Q(\theta|\theta^t) = E_{Z|X,\theta^t} [\log p(X, Z|\theta)]$$  \hspace{1cm} (4)

3) M step: maximize the Q function:

$$\theta^{(t+1)} = \arg\max_{\theta}Q(\theta|\theta^t)$$  \hspace{1cm} (5)

4) Assess change in the log likelihood, iterate 2-3 as necessary
EM Summary

- Expectation-Maximization is a very general algorithm that can solve many optimization problems.
EM Summary

- Expectation-Maximization is a very general algorithm that can solve many optimization problems.
- Works with missing data or latent variables.
EM Summary

- Expectation-Maximization is a very general algorithm that can solve many optimization problems
- Works with missing data or latent variables
- Will play a key role in discussion of missing data
Expectation-Maximization is a very general algorithm that can solve many optimization problems. It works with missing data or latent variables. It will play a key role in discussion of missing data, and there are many variants for dealing with complicated Q functions etc.
EM Summary

- Expectation-Maximization is a very general algorithm that can solve many optimization problems
- Works with missing data or latent variables
- Will play a key role in discussion of missing data
- Many variants for dealing with complicated Q functions etc.
- Related to many approaches in Bayesian computing.
1. Mixture Models
   - Basic Mixtures
   - Application: Mixtures as Preprocessing
   - Application: Mixture of Regressions

2. Expectation Maximization
   - EM for Probit Regression
   - EM for Gaussian Mixtures
   - EM in General

3. Missing Data
   - Motivating Example
   - Overview and Assumptions
   - Existing Heuristics
   - Application Specific Approaches
   - Multiple Imputation
   - The Full Amelia Scheme

4. Measurement Error

5. Appendix: Additional Details and Examples
1 Mixture Models
   - Basic Mixtures
   - Application: Mixtures as Preprocessing
   - Application: Mixture of Regressions

2 Expectation Maximization
   - EM for Probit Regression
   - EM for Gaussian Mixtures
   - EM in General

3 Missing Data
   - Motivating Example
   - Overview and Assumptions
   - Existing Heuristics
   - Application Specific Approaches
   - Multiple Imputation
   - The Full Amelia Scheme

4 Measurement Error

5 Appendix: Additional Details and Examples
The Slovenian Plebiscite (Rubin, Stern and Vehovar, 1995)

In 1990, the Government of Slovenia (at that point, one of several republics within Yugoslavia) administered a poll to determine the extent of support for an upcoming plebiscite on Slovenian independence.
The Slovenian Plebiscite (Rubin, Stern and Vehovar, 1995)

In 1990, the Government of Slovenia (at that point, one of several republics within Yugoslavia) administered a poll to determine the extent of support for an upcoming plebiscite on Slovenian independence. Passage of the plebiscite required that at least 50% of eligible Slovenian voters both turn out and vote for independence.
In 1990, the Government of Slovenia (at that point, one of several republics within Yugoslavia) administered a poll to determine the extent of support for an upcoming plebiscite on Slovenian independence. Passage of the plebiscite required that at least 50% of eligible Slovenian voters both turn out and vote for independence.

Here are the survey results ($n = 2074$):
In 1990, the Government of Slovenia (at that point, one of several republics within Yugoslavia) administered a poll to determine the extent of support for an upcoming plebiscite on Slovenian independence. Passage of the plebiscite required that at least 50% of eligible Slovenian voters both turn out and vote for independence.

Here are the survey results ($n = 2074$):

<table>
<thead>
<tr>
<th>Attendance</th>
<th>Independence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
<td>1439</td>
</tr>
<tr>
<td>No</td>
<td>16</td>
</tr>
<tr>
<td>DK</td>
<td>144</td>
</tr>
</tbody>
</table>
Quantities of Interest

We might assume that all of the “don’t know” folks do in fact have some intentions. We are interested in the proportion of the population in each of the four groups.
Quantities of Interest

We might assume that all of the “don’t know” folks do in fact have some intentions. We are interested in the proportion of the population in each of the four groups.

<table>
<thead>
<tr>
<th>Attendance</th>
<th>Independence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>( \theta_{11} )</td>
</tr>
<tr>
<td>No</td>
<td>( \theta_{21} )</td>
</tr>
</tbody>
</table>

Here the first subscript refers to the attendance question and the second to the independence question.
Some Possible Estimates

Our quantity of interest is the proportion of individuals in the population who both support independence and will attend the plebiscite.
Some Possible Estimates

Our quantity of interest is the proportion of individuals in the population who both support independence and will attend the plebiscite. There are a few possible estimators:

1. Deletion estimator: the proportion is \( \hat{\theta}_{11} = \frac{1439}{1439 + 78 + 16 + 16} = 0.929 \).

2. Conservative estimator: assume that people answering "don't know" are simply trying to avoid revealing an unpopular opinion, so \( \hat{\theta}_{11} = \frac{1439}{1549 + 525} = 0.46938 \).

3. Make some other set of behavioral assumptions about the different missingness blocs.

4. Imputation estimator: assert that the missingness is determined only by the observed values and then attempt to impute the missing data.
Some Possible Estimates

Our quantity of interest is the proportion of individuals in the population who both support independence and will attend the plebiscite. There are a few possible estimators:

1. **Deletion estimator:**

\[
\hat{\theta}_{\text{Deletion}} = \frac{1439}{1439 + 78 + 16 + 16} = 0.929.
\]

Strongly assume that people who “don’t know” will change their preferences to reflect those who do.

2. **Conservative estimator:**

\[
\hat{\theta}_{\text{Conservative}} = \frac{1439}{1549 + 525} = 0.71838.
\]

Assume that people answering “don’t know” are simply trying to avoid revealing an unpopular opinion, so:

3. Make some other set of behavioral assumptions about the different missingness blocs.

4. **Imputation estimator:**

Assert that the missingness is determined only by the observed values and then attempt to impute the missing data.
Some Possible Estimates

Our quantity of interest is the proportion of individuals in the population who both support independence and will attend the plebiscite. There are a few possible estimators:

1. **Deletion estimator**: the proportion is \( \hat{\theta}_{11} = \frac{1439}{1439 + 78 + 16 + 16} = .929 \). Strongly assume that people who “don’t know” will change their preferences to reflect those who do.

2. **Conservative estimator**: assume that people answering “don’t know” are simply trying to avoid revealing an unpopular opinion, so \( \hat{\theta}_{11} = \frac{1439}{1549 + 525} = .6938 \).

3. Make some other set of behavioral assumptions about the different missingness blocs.

4. **Imputation estimator**: assert that the missingness is determined only by the observed values and then attempt to impute the missing data.
Some Possible Estimates

Our quantity of interest is the proportion of individuals in the population who both support independence and will attend the plebiscite. There are a few possible estimators:

1. **Deletion estimator**: the proportion is \( \hat{\theta}_{11} = \frac{1439}{1439 + 78 + 16 + 16} = 0.929. \) Strongly assume that people who “don’t know” will change their preferences to reflect those who do.

2. **Conservative estimator**:
Some Possible Estimates

Our quantity of interest is the proportion of individuals in the population who both support independence and will attend the plebiscite. There are a few possible estimators:

1. **Deletion estimator**: the proportion is \( \hat{\theta}_{11} = \frac{1439}{1439 + 78 + 16 + 16} = 0.929 \). Strongly assume that people who “don’t know” will change their preferences to reflect those who do.

2. **Conservative estimator**: assume that people answering “don’t know” are simply trying to avoid revealing an unpopular opinion, so \( \hat{\theta}_{11} = \frac{1439}{1549 + 525} = 0.6938 \).
Some Possible Estimates

Our quantity of interest is the proportion of individuals in the population who both support independence and will attend the plebiscite. There are a few possible estimators:

1. **Deletion estimator:** the proportion is \( \hat{\theta}_{11} = \frac{1439}{1439 + 78 + 16 + 16} = 0.929 \).

   Strongly assume that people who “don’t know” will change their preferences to reflect those who do.

2. **Conservative estimator:** assume that people answering “don’t know” are simply trying to avoid revealing an unpopular opinion, so \( \hat{\theta}_{11} = \frac{1439}{1549 + 525} = 0.6938 \).

3. Make some other set of behavioral assumptions about the different missingness blocs.
Some Possible Estimates

Our quantity of interest is the proportion of individuals in the population who both support independence and will attend the plebiscite. There are a few possible estimators:

1. **Deletion estimator**: the proportion is \( \hat{\theta}_{11} = \frac{1439}{1439 + 78 + 16 + 16} = .929 \). Strongly assume that people who “don’t know” will change their preferences to reflect those who do.

2. **Conservative estimator**: assume that people answering “don’t know” are simply trying to avoid revealing an unpopular opinion, so \( \hat{\theta}_{11} = \frac{1439}{1549 + 525} = .6938 \).

3. Make some other set of behavioral assumptions about the different missingness blocs.

4. **Imputation estimator**: 

Stewart (Princeton)
Some Possible Estimates

Our quantity of interest is the proportion of individuals in the population who both support independence and will attend the plebiscite. There are a few possible estimators:

1. **Deletion estimator**: the proportion is \( \hat{\theta}_{11} = \frac{1439}{1439 + 78 + 16 + 16} = .929 \). Strongly assume that people who “don’t know” will change their preferences to reflect those who do.

2. **Conservative estimator**: assume that people answering “don’t know” are simply trying to avoid revealing an unpopular opinion, so \( \hat{\theta}_{11} = \frac{1439}{1549 + 525} = .6938 \).

3. Make some other set of behavioral assumptions about the different missingness blocs.

4. **Imputation estimator**: assert that the missingness is determined only by the observed values and then attempt to impute the missing data.
Here’s the data again, with the proportion of observed data filled in.

<table>
<thead>
<tr>
<th>Attendance</th>
<th>Independence</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>DK</td>
</tr>
<tr>
<td>Yes</td>
<td>1439 (.928)</td>
<td>78 (.050)</td>
<td>159</td>
</tr>
<tr>
<td>No</td>
<td>16 (.010)</td>
<td>16 (.010)</td>
<td>32</td>
</tr>
<tr>
<td>DK</td>
<td>144</td>
<td>54</td>
<td>136</td>
</tr>
</tbody>
</table>
Imputation

Well, among fully observed individuals we can see that \( \frac{.928}{.928 + .050} = .949 \) of the A-Y folks will vote I-Y.
Imputation

Well, among fully observed individuals we can see that $\frac{.928}{.928+.050} = .949$ of the A-Y folks will vote I-Y. So we might guess the same for those who didn’t answer the independence question.
Well, among fully observed individuals we can see that $$\frac{.928}{.928+.050} = .949$$ of the A-Y folks will vote I-Y. So we might guess the same for those who didn’t answer the independence question.

Well, among fully observed individuals we can see that \( \frac{.928}{.928 + .050} = .949 \) of the A-Y folks will vote I-Y. So we might guess the same for those who didn’t answer the independence question.

\[
\]

This means that the expected number of I-N votes among A-Y, I-DK is now \( 159 - 150.87 = 8.13 \).
Imputation

Well, among fully observed individuals we can see that \( \frac{.928}{.928 + .050} = .949 \) of the A-Y folks will vote I-Y. So we might guess the same for those who didn’t answer the independence question.

\[
E[A - Y, I - Y'\text{'s among } A - Y, I - DK'\text{'s}] = 159 \times .949 = 150.87.
\]

This means that the expected number of I-N votes among A-Y,I-DK is now \( 159 - 150.87 = 8.13 \).

We can do exactly the same set of calculations for the other three “don’t know” groups to impute the missing data.
We have made a guess of missing values based on estimates of population parameters $\theta$. What would be a suitable next step?

**Table: Imputations for I-DK’s in red; imputations based on A-DK’s in blue.**

<table>
<thead>
<tr>
<th>Attendance</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>1439 + 150.87 + 142.42</td>
<td>78 + 8.12 + 44.81</td>
</tr>
<tr>
<td></td>
<td>.896</td>
<td>.066</td>
</tr>
<tr>
<td>No</td>
<td>16 + 16 + 1.58</td>
<td>16 + 16 + 9.19</td>
</tr>
<tr>
<td></td>
<td>.017</td>
<td>.020</td>
</tr>
</tbody>
</table>
Imputation: An Updated Sense of the Proportions?

<table>
<thead>
<tr>
<th>Attendance</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>1439 + 150.87 + 142.42</td>
<td>78 + 8.12 + 44.81</td>
</tr>
<tr>
<td></td>
<td>.896</td>
<td>.066</td>
</tr>
<tr>
<td>No</td>
<td>16 + 16 + 1.58</td>
<td>16 + 16 + 9.19</td>
</tr>
<tr>
<td></td>
<td>.017</td>
<td>.020</td>
</tr>
</tbody>
</table>

Table: Imputations for I-DK’s in red; imputations based on A-DK’s in blue.

We have made a guess of missing values based on estimates of population parameters $\theta$. What would be a suitable next step?
We can now use our updated (and, in fact, improved) estimate of the population proportions in order to re-impute the missing data using the same approach as before.
Iteration

We can now use our updated (and, in fact, improved) estimate of the population proportions in order to re-impute the missing data using the same approach as before.

Once we have updated our best guess of how the various DK people will vote, then we can re-estimate the population proportions.
Iteration

We can now use our updated (and, in fact, improved) estimate of the population proportions in order to re-impute the missing data using the same approach as before.

Once we have updated our best guess of how the various DK people will vote, then we can re-estimate the population proportions.

We can iterate this approach until our estimates of the population proportions converge to a stable maximum.
Iterations

Here are the trace plots showing how the estimates of the $\theta$ evolve through the iterations:
A Final Estimate

After running the algorithm for 30 iterations, the final estimate for $\theta_{11}$ was $\hat{\theta}_{11} = .892$. Two weeks after this survey was conducted the plebiscite was held, and it turned out that 88.5% of eligible voters turned out and voted for independence.
A Final Estimate

After running the algorithm for 30 iterations, the final estimate for $\theta_{11}$ was $\hat{\theta}_{11} = .892$.

Recall that our original deletion estimator estimate was .928.
A Final Estimate

After running the algorithm for 30 iterations, the final estimate for $\theta_{11}$ was

$$\hat{\theta}_{11} = .892.$$  

Recall that our original deletion estimator estimate was .928.

Two weeks after this survey was conducted the plebiscite was held, and it turned out that 88.5% of eligible voters turned out and voted for independence.
A Final Estimate

After running the algorithm for 30 iterations, the final estimate for $\theta_{11}$ was $\hat{\theta}_{11} = .892$.

Recall that our original deletion estimator estimate was .928.

Two weeks after this survey was conducted the plebiscite was held, and it turned out that 88.5% of eligible voters turned out and voted for independence.

Neat!
1 Mixture Models
   - Basic Mixtures
   - Application: Mixtures as Preprocessing
   - Application: Mixture of Regressions

2 Expectation Maximization
   - EM for Probit Regression
   - EM for Gaussian Mixtures
   - EM in General

3 Missing Data
   - Motivating Example
   - Overview and Assumptions
   - Existing Heuristics
   - Application Specific Approaches
   - Multiple Imputation
   - The Full Amelia Scheme

4 Measurement Error

5 Appendix: Additional Details and Examples
Missing Data Overview

Missing data is a common problem in applied work. Most of the solutions will turn on assumptions about the mechanism that drives the missingness, much as our discussion of causal inference turned on our ability to describe the assignment mechanism.

There are many biased or inefficient missing data practices:

- making up numbers e.g. changing an opinion question to "don't know"
- listwise deletion e.g. most widely used and statistical software default
- indicator variables e.g. including a dummy variable for missing observations
- many other ad hoc approaches

There are three general approaches:

- Imputation: methods for filling in values
- Sensitivity: tests for variation in results
- Bounds: determining the range of possible values under different missingness strategies

We will (mostly) focus on imputation.
Missing Data Overview

- Missing data is a common problem in applied work
Missing Data Overview

- Missing data is a common problem in applied work
- Most of the solutions will turn on assumptions about the mechanism that drives the missingness, much as our discussion of causal inference turned on our ability to describe the assignment mechanism
Missing Data Overview

- Missing data is a common problem in applied work
- Most of the solutions will turn on assumptions about the mechanism that drives the missingness, much as our discussion of causal inference turned on our ability to describe the assignment mechanism
- There are many biased or inefficient missing data practices:
Missing Data Overview

- Missing data is a common problem in applied work
- Most of the solutions will turn on assumptions about the mechanism that drives the missingness, much as our discussion of causal inference turned on our ability to describe the assignment mechanism.
- There are many biased or inefficient missing data practices:
  - making up numbers e.g. changing an opinion question to “don’t know”
Missing Data Overview

- Missing data is a common problem in applied work.
- Most of the solutions will turn on assumptions about the mechanism that drives the missingness, much as our discussion of causal inference turned on our ability to describe the assignment mechanism.
- There are many biased or inefficient missing data practices:
  - making up numbers e.g. changing an opinion question to “don’t know”
  - listwise deletion e.g. most widely used and statistical software default
Missing Data Overview

- Missing data is a common problem in applied work
- Most of the solutions will turn on assumptions about the mechanism that drives the missingness, much as our discussion of causal inference turned on our ability to describe the assignment mechanism
- There are many biased or inefficient missing data practices:
  - making up numbers e.g. changing an opinion question to “don’t know”
  - listwise deletion e.g. most widely used and statistical software default
  - indicator variables e.g. including a dummy variable for missing observations
Missing Data Overview

- Missing data is a common problem in applied work.
- Most of the solutions will turn on assumptions about the mechanism that drives the missingness, much as our discussion of causal inference turned on our ability to describe the assignment mechanism.
- There are many biased or inefficient missing data practices:
  - making up numbers e.g. changing an opinion question to “don’t know”
  - listwise deletion e.g. most widely used and statistical software default
  - indicator variables e.g. including a dummy variable for missing observations
  - many other ad hoc approaches

There are three general approaches:
- Imputation: methods for filling in values
- Sensitivity: tests for variation in results
- Bounds: determining the range of possible values under different missingness strategies

We will (mostly) focus on imputation.
Missing Data Overview

- Missing data is a common problem in applied work
- Most of the solutions will turn on assumptions about the mechanism that drives the missingness, much as our discussion of causal inference turned on our ability to describe the assignment mechanism
- There are many biased or inefficient missing data practices:
  - making up numbers e.g. changing an opinion question to “don’t know”
  - listwise deletion e.g. most widely used and statistical software default
  - indicator variables e.g. including a dummy variable for missing observations
  - many other ad hoc approaches
- There are three general approaches:
Missing Data Overview

- Missing data is a common problem in applied work
- Most of the solutions will turn on assumptions about the mechanism that drives the missingness, much as our discussion of causal inference turned on our ability to describe the assignment mechanism
- There are many biased or inefficient missing data practices:
  - making up numbers e.g. changing an opinion question to “don’t know”
  - listwise deletion e.g. most widely used and statistical software default
  - indicator variables e.g. including a dummy variable for missing observations
  - many other ad hoc approaches
- There are three general approaches:
  - Imputation: methods for filling in values
Missing Data Overview

- Missing data is a common problem in applied work
- Most of the solutions will turn on assumptions about the mechanism that drives the missingness, much as our discussion of causal inference turned on our ability to describe the assignment mechanism
- There are many biased or inefficient missing data practices:
  - making up numbers e.g. changing an opinion question to “don’t know”
  - listwise deletion e.g. most widely used and statistical software default
  - indicator variables e.g. including a dummy variable for missing observations
  - many other ad hoc approaches
- There are three general approaches:
  - Imputation: methods for filling in values
  - Sensitivity: tests for variation in results
Missing Data Overview

- Missing data is a common problem in applied work
- Most of the solutions will turn on assumptions about the mechanism that drives the missingness, much as our discussion of causal inference turned on our ability to describe the assignment mechanism
- There are many biased or inefficient missing data practices:
  - making up numbers e.g. changing an opinion question to “don’t know”
  - listwise deletion e.g. most widely used and statistical software default
  - indicator variables e.g. including a dummy variable for missing observations
  - many other ad hoc approaches
- There are three general approaches:
  - Imputation: methods for filling in values
  - Sensitivity: tests for variation in results
  - Bounds: determining the range of possible values under different missingness strategies
Missing Data Overview

- Missing data is a common problem in applied work
- Most of the solutions will turn on assumptions about the mechanism that drives the missingness, much as our discussion of causal inference turned on our ability to describe the assignment mechanism
- There are many biased or inefficient missing data practices:
  - making up numbers e.g. changing an opinion question to “don’t know”
  - listwise deletion e.g. most widely used and statistical software default
  - indicator variables e.g. including a dummy variable for missing observations
  - many other ad hoc approaches
- There are three general approaches:
  - Imputation: methods for filling in values
  - Sensitivity: tests for variation in results
  - Bounds: determining the range of possible values under different missingness strategies
- We will (mostly) focus on imputation
The Goal of Missing Data Analysis is Population Inference

Missing data is a nuisance for applied work and it is easy to lose sight of the ultimate goal. We want to make a population inference, not to estimate, predict or recover missing observations. Even though we may occasionally check our procedures this way, our goal isn't really to reproduce the results of the complete data analysis. Mean imputation (replacing missing data with the population mean) may be reasonably predictive of the missing data by some metric, but it distorts the variances and covariances which are key to inference. In this sense, we cannot really separate the missing data procedure from the inferential goal of the analysis.
Missing data is a nuisance for applied work and it is easy to lose sight of the ultimate goal.
The Goal of Missing Data Analysis is Population Inference

- Missing data is a nuisance for applied work and it is easy to lose sight of the ultimate goal.
- We want to make a population inference, not to estimate, predict or recover missing observations.
The Goal of Missing Data Analysis is Population Inference

- Missing data is a nuisance for applied work and it is easy to lose sight of the ultimate goal.
- We want to make a population inference, not to estimate, predict or recover missing observations.
- Even though we may occasionally check our procedures this way, our goal isn’t really to reproduce the results of the complete data analysis.
The Goal of Missing Data Analysis is Population Inference

- Missing data is a nuisance for applied work and it is easy to lose sight of the ultimate **goal**.
- We want to make a **population inference**, not to **estimate**, **predict** or **recover** missing observations.
- Even though we may occasionally check our procedures this way, our goal isn’t really to reproduce the results of the complete data analysis.
- **Mean imputation** (replacing missing data with the population mean) may be reasonably predictive of the missing data by some metric, but it distorts the variances and covariances which are key to inference.
The Goal of Missing Data Analysis is Population Inference

- Missing data is a nuisance for applied work and it is easy to lose sight of the ultimate goal.
- We want to make a population inference, not to estimate, predict or recover missing observations.
- Even though we may occasionally check our procedures this way, our goal isn’t really to reproduce the results of the complete data analysis.
- **Mean imputation** (replacing missing data with the population mean) may be reasonably predictive of the missing data by some metric, but it distorts the variances and covariances which are key to inference.
- In this sense- we cannot really separate the missing data procedure from the inferential goal of the analysis.
Missingness Notation

\[ D = \begin{bmatrix} 1 & 2.5 & 432 & 0.5 & 32 & 5 \ 1.2 & 543 & 12 & 7 & 421 & 9 \ 6 & 1.9 & 234 & 1 & 3.2 & 108 \ 0 & 7 & 95 & 1 \end{bmatrix}, \]

\[ M = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \]

\[ D_{\text{mis}} = \text{missing elements in } D \text{ (in Red)} \]

\[ D_{\text{obs}} = \text{observed elements in } D; \]

Missing elements must exist (what's your view on the National Helium Reserve?)

Stewart (Princeton)
Missingness Notation

\[
D = \begin{pmatrix}
1 & 2.5 & 432 & 0 \\
5 & 3.2 & 543 & 1 \\
2 & 7.4 & 219 & 1 \\
6 & 1.9 & 234 & 1 \\
3 & 1.2 & 108 & 0 \\
0 & 7.7 & 95 & 1
\end{pmatrix}, \quad M = \begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 \\
0 & 1 & 1 & 1 \\
0 & 1 & 1 & 1
\end{pmatrix}
\]
### Missingness Notation

$$D = \begin{pmatrix}
1 & 2.5 & 432 & 0 \\
5 & 3.2 & 543 & 1 \\
2 & 7.4 & 219 & 1 \\
6 & 1.9 & 234 & 1 \\
3 & 1.2 & 108 & 0 \\
0 & 7.7 & 95 & 1
\end{pmatrix}, \quad M = \begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 \\
0 & 1 & 1 & 1 \\
0 & 1 & 1 & 1
\end{pmatrix}$$

\(D_{mis} = \text{missing} \) elements in \(D\) (in Red)
Missingness Notation

\[ D = \begin{pmatrix}
1 & 2.5 & 432 & 0 \\
5 & 3.2 & 543 & 1 \\
2 & 7.4 & 219 & 1 \\
6 & 1.9 & 234 & 1 \\
3 & 1.2 & 108 & 0 \\
0 & 7.7 & 95 & 1 \\
\end{pmatrix}, \quad M = \begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 \\
0 & 1 & 1 & 1 \\
0 & 1 & 1 & 1 \\
\end{pmatrix} \]

\( D_{mis} = \text{missing elements in } D \text{ (in Red)} \)

\( D_{obs} = \text{observed elements in } D \)
Missingness Notation

\[ D = \begin{pmatrix} 1 & 2.5 & 432 & 0 \\ 5 & 3.2 & 543 & 1 \\ 2 & 7.4 & 219 & 1 \\ 6 & 1.9 & 234 & 1 \\ 3 & 1.2 & 108 & 0 \\ 0 & 7.7 & 95 & 1 \end{pmatrix}, \quad M = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix} \]

\[ D_{mis} = \text{missing} \text{ elements in } D \text{ (in Red)} \]
\[ D_{obs} = \text{observed elements in } D \]

\implies \text{Missing elements must exist (what's your view on the National Helium Reserve?)}
What Can Be Learned With Minimal Assumptions

We will introduce a powerful set of assumptions which suggest alternate strategies. To motivate these assumptions, let’s consider what can be learned with very few assumptions using the framework of sharp Manski bounds (see e.g. Aronow and Miller Chapter 4).

Assumptions:

\[ Y_i \text{ is bounded with support } [a, b] \text{ and we assume stable outcomes } \]

\[ Y^*_i = Y_i + (NA)(1 - M_i) \]

which simply suggests that the \( Y_i \) is stable (e.g. regardless of how the question is asked or who responded).

We obtain sharp bounds for \( E[Y] \) by first plugging in \( a \) for all missing values to get the lower bound, followed by plugging in \( b \) for all missing values to get the upper bound. This leaves our quantity set identified as opposed to our usual point identified, without further assumptions we can do no better. This only works with bounded support and becomes much harder with missingness on many variables.
What Can Be Learned With Minimal Assumptions

- We will introduce a powerful set of assumptions which suggest alternate strategies.

Assumptions:

\[ Y_i \] is bounded with support \([a, b]\) and we assume stable outcomes \(Y^*_i = Y_i M_i (1 - M_i)\) which simply suggests that the \(Y_i\) is stable (e.g. regardless of how the question is asked or who responded).

We obtain sharp bounds for \(E[Y]\) by first plugging in \(a\) for all missing values to get the lower bound, followed by plugging in \(b\) for all missing values to get the upper bound.

This leaves our quantity set identified as opposed to our usual point identified, without further assumptions we can do no better.

This only works with bounded support and becomes much harder with missingness on many variables.
What Can Be Learned With Minimal Assumptions

- We will introduce a powerful set of assumptions which suggest alternate strategies.
- To motivate these assumptions, let’s consider what can be learned with very few assumptions using the framework of sharp Manski bounds (see e.g. Aronow and Miller Chapter 4).
What Can Be Learned With Minimal Assumptions

- We will introduce a powerful set of assumptions which suggest alternate strategies.
- To motivate these assumptions, let’s consider what can be learned with very few assumptions using the framework of sharp Manski bounds (see e.g. Aronow and Miller Chapter 4).
- Assumptions: $Y_i$ is bounded with support $[a, b]$ and we assume stable outcomes $Y_i^* = Y_iM_i + (NA)(1 - M_i)$ which simply suggests that the $Y_i$ is stable (e.g. regardless of how the question is asked or who responded).
What Can Be Learned With Minimal Assumptions

- We will introduce a powerful set of assumptions which suggest alternate strategies.
- To motivate these assumptions, let’s consider what can be learned with very few assumptions using the framework of sharp Manski bounds (see e.g. Aronow and Miller Chapter 4).
- Assumptions: \( Y_i \) is bounded with support \([a, b]\) and we assume stable outcomes \( Y_i^* = Y_i M_i + (\text{NA})(1 - M_i) \) which simply suggests that the \( Y_i \) is stable (e.g. regardless of how the question is asked or who responded).
- We obtain sharp bounds for \( E[Y] \) by first plugging in \( a \) for all missing values to get the lower bound, followed by plugging in \( b \) for all missing values to get the upper bound.
What Can Be Learned With Minimal Assumptions

- We will introduce a powerful set of assumptions which suggest alternate strategies.
- To motivate these assumptions, let’s consider what can be learned with very few assumptions using the framework of sharp Manski bounds (see e.g. Aronow and Miller Chapter 4).
- Assumptions: $Y_i$ is bounded with support $[a, b]$ and we assume stable outcomes $Y_i^* = Y_i M_i + (\text{NA})(1 - M_i)$ which simply suggests that the $Y_i$ is stable (e.g. regardless of how the question is asked or who responded).
- We obtain sharp bounds for $E[Y]$ by first plugging in $a$ for all missing values to get the lower bound, followed by plugging in $b$ for all missing values to get the upper bound.
- This leaves our quantity set identified as opposed to our usual point identified, without further assumptions we can do no better.
What Can Be Learned With Minimal Assumptions

- We will introduce a powerful set of assumptions which suggest alternate strategies.
- To motivate these assumptions, let’s consider what can be learned with very few assumptions using the framework of sharp Manski bounds (see e.g. Aronow and Miller Chapter 4).
- Assumptions: \( Y_i \) is bounded with support \([a, b]\) and we assume stable outcomes \( Y_i^* = Y_i M_i + (NA)(1 - M_i) \) which simply suggests that the \( Y_i \) is stable (e.g. regardless of how the question is asked or who responded).
- We obtain sharp bounds for \( E[Y] \) by first plugging in \( a \) for all missing values to get the lower bound, followed by plugging in \( b \) for all missing values to get the upper bound.
- This leaves our quantity set identified as opposed to our usual point identified, without further assumptions we can do no better.
- This only works with bounded support and becomes much harder with missingness on many variables.
Possible Further Assumptions

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Acronym</th>
<th>You can predict ( M ) with:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Completely At Random</td>
<td>MCAR</td>
<td></td>
</tr>
<tr>
<td>Missing At Random</td>
<td>MAR</td>
<td></td>
</tr>
<tr>
<td>Nonignorable</td>
<td>NI</td>
<td></td>
</tr>
</tbody>
</table>

Reasons for the odd terminology are historical.

Stewart (Princeton)
Possible Further Assumptions

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Acronym</th>
<th>You can predict $M$ with:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Completely At Random</td>
<td>MCAR</td>
<td></td>
</tr>
<tr>
<td>Missing At Random</td>
<td>MAR</td>
<td></td>
</tr>
<tr>
<td>Nonignorable</td>
<td>NI</td>
<td></td>
</tr>
<tr>
<td>&amp; Obs &amp; Mis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Reasons for the odd terminology are historical.
## Possible Further Assumptions

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Acronym</th>
<th>You can predict $M$ with:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Completely At Random</td>
<td>MCAR</td>
<td>—</td>
</tr>
</tbody>
</table>

Reasons for the odd terminology are historical.
## Possible Further Assumptions

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Acronym</th>
<th>You can predict $M$ with:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Completely At Random</td>
<td>MCAR</td>
<td>—</td>
</tr>
<tr>
<td>Missing At Random</td>
<td>MAR</td>
<td>$D_{obs}$</td>
</tr>
</tbody>
</table>
## Possible Further Assumptions

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Acronym</th>
<th>You can predict $M$ with:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Completely At Random</td>
<td>MCAR</td>
<td>—</td>
</tr>
<tr>
<td>Missing At Random</td>
<td>MAR</td>
<td>$D_{obs}$</td>
</tr>
<tr>
<td>Nonignorable</td>
<td>NI</td>
<td>$D_{obs}$ &amp; $D_{mis}$</td>
</tr>
</tbody>
</table>
Possible Further Assumptions

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Acronym</th>
<th>You can predict $M$ with:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Completely At Random</td>
<td>MCAR</td>
<td>—</td>
</tr>
<tr>
<td>Missing At Random</td>
<td>MAR</td>
<td>$D_{obs}$</td>
</tr>
<tr>
<td>Nonignorable</td>
<td>NI</td>
<td>$D_{obs} &amp; D_{mis}$</td>
</tr>
</tbody>
</table>

- Reasons for the odd terminology are historical.
Missingness Assumptions, again

1. MCAR: Coin flips determine whether to answer survey questions (naive)

   \[ P(M|D) = P(M) \]

2. MAR: missingness is a function of measured variables (empirical)

   \[ P(M|D) \equiv P(M|D_{\text{obs}}, D_{\text{mis}}) = P(M|D_{\text{obs}}) \]

   ▶ e.g., Independents are less likely to answer vote choice question (with PID measured)
   ▶ e.g., Some occupations are less likely to answer the income question (with occupation measured)

3. NI: missingness depends on unobservables (fatalistic)

   \[ P(M|D) \text{ doesn't simplify} \]
   ▶ e.g., censoring income if income is $100K and you can't predict high income with other measured variables
   ▶ Adding variables to predict income can change NI to MAR
Missingness Assumptions, again

1. **MCAR**: Coin flips determine whether to answer survey questions (naive)

   - e.g., Independents are less likely to answer vote choice question (with PID measured)
   - e.g., Some occupations are less likely to answer the income question (with occupation measured)

2. **MAR**: missingness is a function of measured variables (empirical)

   - e.g., censoring income if income is >$100K and you can't predict high income with other measured variables
   - Adding variables to predict income can change NI to MAR

3. **NI**: missingness depends on unobservables (fatalistic)

   - $P(M|D)$ doesn't simplify
   - e.g., censoring income if income is >$100K and you can't predict high income with other measured variables

   - Adding variables to predict income can change NI to MAR
Missingness Assumptions, again

1. **MCAR**: Coin flips determine whether to answer survey questions (naive)

\[ P(M|D) = P(M) \]

2. **MAR**: missingness is a function of measured variables (empirical)

\[ P(M|D) \equiv P(M|D_{\text{obs}}, D_{\text{mis}}) = P(M|D_{\text{obs}}) \]

- e.g., Independents are less likely to answer vote choice question (with PID measured)
- e.g., Some occupations are less likely to answer the income question (with occupation measured)

3. **NI**: missingness depends on unobservables (fatalistic)

\[ P(M|D) \text{ doesn't simplify} \]

- e.g., censoring income if income is > $100K and you can't predict high income with other measured variables
- Adding variables to predict income can change NI to MAR
Missingness Assumptions, again

1. **MCAR**: Coin flips determine whether to answer survey questions (naive)

   $$\mathbb{P}(M|D) = \mathbb{P}(M)$$

2. **MAR**: missingness is a function of measured variables (empirical)

   ▶ e.g., Independents are less likely to answer vote choice question (with PID measured)
   ▶ e.g., Some occupations are less likely to answer the income question (with occupation measured)

3. **NI**: missingness depends on unobservables (fatalistic)

   ▶ $$\mathbb{P}(M|D)$$ doesn’t simplify
   ▶ e.g., censoring income if income is $\text{\$100K}$ and you can’t predict high income with other measured variables
   ▶ Adding variables to predict income can change NI to MAR
Missingness Assumptions, again

1. **MCAR**: Coin flips determine whether to answer survey questions (**naive**)

   \[ P(M|D) = P(M) \]

2. **MAR**: missingness is a function of measured variables (**empirical**)

   \[ P(M|D) \equiv P(M|D_{obs}, D_{mis}) = P(M|D_{obs}) \]

   - e.g., Independents are less likely to answer vote choice question (with PID measured)
   - e.g., Some occupations are less likely to answer the income question (with occupation measured)

3. **NI**: missingness depends on unobservables (**fatalistic**)

   - e.g., censoring income if income is $100K and you can't predict high income with other measured variables
   - Adding variables to predict income can change NI to MAR
Missingness Assumptions, again

1. **MCAR**: Coin flips determine whether to answer survey questions (*naive*)

   \[ \mathbb{P}(M \mid D) = \mathbb{P}(M) \]

2. **MAR**: missingness is a function of measured variables (*empirical*)

   \[ \mathbb{P}(M \mid D) \equiv \mathbb{P}(M \mid D_{obs}, D_{mis}) = \mathbb{P}(M \mid D_{obs}) \]

   - e.g., Independents are less likely to answer vote choice question (with PID measured)

   - e.g., Some occupations are less likely to answer the income question (with occupation measured)

3. **NI**: missingness depends on unobservables (*fatalistic*)

   \[ \mathbb{P}(M \mid D) \text{ doesn't simplify} \]

   - e.g., censoring income if income is $100K and you can't predict high income with other measured variables

   - Adding variables to predict income can change NI to MAR
Missingness Assumptions, again

1. **MCAR**: Coin flips determine whether to answer survey questions (naive)

   \[
P(M|D) = P(M)
   \]

2. **MAR**: missingness is a function of measured variables (empirical)

   \[
P(M|D) \equiv P(M|D_{obs}, D_{mis}) = P(M|D_{obs})
   \]

   ▶ e.g., Independents are less likely to answer vote choice question (with PID measured)
   ▶ e.g., Some occupations are less likely to answer the income question (with occupation measured)
Missingness Assumptions, again

1. **MCAR**: Coin flips determine whether to answer survey questions (*naive*)

   \[ P(M|D) = P(M) \]

2. **MAR**: missingness is a function of measured variables (*empirical*)

   \[ P(M|D) \equiv P(M|D_{obs}, D_{mis}) = P(M|D_{obs}) \]

   - e.g., Independents are less likely to answer vote choice question (with PID measured)
   - e.g., Some occupations are less likely to answer the income question (with occupation measured)

3. **NI**: missingness depends on unobservables (*fatalistic*)

   \[ P(M|D) \text{ doesn't simplify} \]
   
   - e.g., censoring income if income is >$100K$ and you can't predict high income with other measured variables
   - Adding variables to predict income can change NI to MAR
Missingness Assumptions, again

1. **MCAR**: Coin flips determine whether to answer survey questions (naive)

\[ P(M|D) = P(M) \]

2. **MAR**: missingness is a function of measured variables (empirical)

\[ P(M|D) \equiv P(M|D_{obs}, D_{mis}) = P(M|D_{obs}) \]

- e.g., Independents are less likely to answer vote choice question (with PID measured)
- e.g., Some occupations are less likely to answer the income question (with occupation measured)

3. **NI**: missingness depends on unobservables (fatalistic)

- \( P(M|D) \) doesn’t simplify
Missingness Assumptions, again

1. **MCAR**: Coin flips determine whether to answer survey questions (naive)
   \[
   \mathbb{P}(M|D) = \mathbb{P}(M)
   \]

2. **MAR**: missingness is a function of measured variables (empirical)
   \[
   \mathbb{P}(M|D) \equiv \mathbb{P}(M|D_{obs}, D_{mis}) = \mathbb{P}(M|D_{obs})
   \]
   ▶ e.g., Independents are less likely to answer vote choice question (with PID measured)
   ▶ e.g., Some occupations are less likely to answer the income question (with occupation measured)

3. **NI**: missingness depends on unobservables (fatalistic)
   ▶ \( P(M|D) \) doesn’t simplify
   ▶ e.g., censoring income if income is $\text{>$100K}$ and you can’t predict high income with other measured variables
Missingness Assumptions, again

1. **MCAR**: Coin flips determine whether to answer survey questions (naive)
   \[ P(M|D) = P(M) \]

2. **MAR**: missingness is a function of measured variables (empirical)
   \[ P(M|D) \equiv P(M|D_{obs}, D_{mis}) = P(M|D_{obs}) \]
   - e.g., Independents are less likely to answer vote choice question (with PID measured)
   - e.g., Some occupations are less likely to answer the income question (with occupation measured)

3. **NI**: missingness depends on unobservables (fatalistic)
   - \( P(M|D) \) doesn’t simplify
   - e.g., censoring income if income is \( > $100K \) and you can’t predict high income with other measured variables
   - Adding variables to predict income can change NI to MAR
1 Mixture Models
   - Basic Mixtures
   - Application: Mixtures as Preprocessing
   - Application: Mixture of Regressions

2 Expectation Maximization
   - EM for Probit Regression
   - EM for Gaussian Mixtures
   - EM in General

3 Missing Data
   - Motivating Example
   - Overview and Assumptions
   - Existing Heuristics
   - Application Specific Approaches
   - Multiple Imputation
   - The Full Amelia Scheme

4 Measurement Error

5 Appendix: Additional Details and Examples
1 Mixture Models
   • Basic Mixtures
   • Application: Mixtures as Preprocessing
   • Application: Mixture of Regressions

2 Expectation Maximization
   • EM for Probit Regression
   • EM for Gaussian Mixtures
   • EM in General

3 Missing Data
   • Motivating Example
   • Overview and Assumptions
   • Existing Heuristics
     • Application Specific Approaches
     • Multiple Imputation
     • The Full Amelia Scheme

4 Measurement Error

5 Appendix: Additional Details and Examples
How Bad is Listwise Deletion?
How Bad is Listwise Deletion?

BAD
How Bad is Listwise Deletion?

BAD

Appendix
Existing General Purpose Missing Data Methods

Fill in or delete the missing data, and then act as if there were no missing data. None work in general under MAR.
Existing General Purpose Missing Data Methods

Fill in or delete the missing data, and then act as if there were no missing data. None work in general under MAR.
Existing General Purpose Missing Data Methods

Fill in or delete the missing data, and then act as if there were no missing data. None work in general under MAR.

Methods which discard data

1. Listwise Deletion (aka complete case)
   empirically RMSE is 1 SE off if MCAR holds; biased under MAR
Existing General Purpose Missing Data Methods

Fill in or delete the missing data, and then act as if there were no missing data. None work in general under MAR.

Methods which discard data

1. Listwise Deletion (aka complete case)
   empirically RMSE is 1 SE off if MCAR holds; biased under MAR

2. Pairwise Deletion
   assumes MCAR; can have numerical stability problems
Existing General Purpose Missing Data Methods

Fill in or delete the missing data, and then act as if there were no missing data. None work in general under MAR.

Methods which discard data

1. Listwise Deletion (aka complete case)
   empirically RMSE is 1 SE off if MCAR holds; biased under MAR

2. Pairwise Deletion
   assumes MCAR; can have numerical stability problems

3. Available Case (aka using only completely observed variables)
   induces omitted variable bias
Existing General Purpose Missing Data Methods

Fill in or delete the missing data, and then act as if there were no missing data. None work in general under MAR.

Methods which discard data

1. Listwise Deletion (aka complete case)  
   empirically RMSE is 1 SE off if MCAR holds; biased under MAR

2. Pairwise Deletion  
   assumes MCAR; can have numerical stability problems

3. Available Case (aka using only completely observed variables)  
   induces omitted variable bias

4. Nonresponse Weighting (including HT weights, Hajek weights)  
   unbiased and consistent but inefficient and high variability in small samples
Simple approaches which retain all the data

5. Mean Imputation
   severely distorts distribution, pulls correlations to zero
Simple approaches which retain all the data

5. Mean Imputation
   severely distorts distribution, pulls correlations to zero

6. Best guess imputation or logical rules
   depends on guesser/logic
Simple approaches which retain all the data

5. Mean Imputation
   severely distorts distribution, pulls correlations to zero

6. Best guess imputation or logical rules
   depends on guesser/logic

7. Indicator for Continuous Variable (aka dummy for missingness)
   biased on other predictors, can include interactions between indicators
   and other predictors which leads to complete case style behavior
Simple approaches which retain all the data

5. Mean Imputation
   severely distorts distribution, pulls correlations to zero

6. Best guess imputation or logical rules
   depends on guesser/logic

7. Indicator for Continuous Variable (aka dummy for missingness)
   biased on other predictors, can include interactions between indicators
   and other predictors which leads to complete case style behavior

8. “Missing” Category for Categorical Variable
   simple and often useful but differential rates in how missingness spreads
   over categories could cause bias
Random Imputation for one variable

9. Simple Random Imputation
   ignores useful information, helpful as a starting point
Random Imputation for one variable

9. Simple Random Imputation
   ignores useful information, helpful as a starting point

10. Hot Deck Imputation (aka matching imputation)
    consistent but otherwise bad: inefficient, standard errors wrong
11. Regression Imputation (aka regression deterministic optimisitic): scatter when observed, perfectly linear when unobserved; SEs too small

12. \( \hat{y} + \epsilon \) regression imputation (aka regression predictive) assumes no estimation uncertainty, does not help for scattered missingness
11. \( \hat{y} \) Regression Imputation (aka regression deterministic) optimistic: scatter when observed, perfectly linear when unobserved; SEs too small
11. \( \hat{y} \) Regression Imputation (aka regression deterministic)
optimistic: scatter when observed, perfectly linear when unobserved;
SEs too small

12. \( \hat{y} + \epsilon \) regression imputation (aka regression predictive)
assumes no estimation uncertainty, does not help for scattered
missingness
1 Mixture Models
   - Basic Mixtures
   - Application: Mixtures as Preprocessing
   - Application: Mixture of Regressions

2 Expectation Maximization
   - EM for Probit Regression
   - EM for Gaussian Mixtures
   - EM in General

3 Missing Data
   - Motivating Example
   - Overview and Assumptions
   - Existing Heuristics
   - Application Specific Approaches
     - Multiple Imputation
     - The Full Amelia Scheme

4 Measurement Error

5 Appendix: Additional Details and Examples
Application-specific Methods for Missing Data

1. Base inferences on the likelihood function or posterior distribution, by conditioning on observed data only, $P(\theta | Y_{obs})$.

2. E.g., models of censoring, truncation, etc.

3. Optimal theoretically, if specification is correct.

4. Not robust (i.e., sensitive to distributional assumptions), a problem if model is incorrect.

5. Often difficult practically.

6. Very difficult with missingness scattered through $X$ and $Y$. 

Stewart (Princeton)
1. Base inferences on the likelihood function or posterior distribution, by conditioning on observed data only, $P(\theta | Y_{obs})$. 

2. E.g., models of censoring, truncation, etc.

3. Optimal theoretically, if specification is correct

4. Not robust (i.e., sensitive to distributional assumptions), a problem if model is incorrect

5. Often difficult practically

6. Very difficult with missingness scattered through $X$ and $Y$.
1. Base inferences on the likelihood function or posterior distribution, by conditioning on observed data only, $P(\theta | Y_{obs})$.

2. E.g., models of censoring, truncation, etc.
1. Base inferences on the likelihood function or posterior distribution, by conditioning on observed data only, $P(\theta|Y_{obs})$.

2. E.g., models of censoring, truncation, etc.

3. Optimal theoretically, if specification is correct.
1. Base inferences on the likelihood function or posterior distribution, by conditioning on observed data only, $P(\theta|Y_{obs})$.

2. E.g., models of censoring, truncation, etc.

3. Optimal theoretically, if specification is correct

4. Not robust (i.e., sensitive to distributional assumptions), a problem if model is incorrect
Application-specific Methods for Missing Data

1. Base inferences on the likelihood function or posterior distribution, by conditioning on observed data only, \( P(\theta | Y_{\text{obs}}) \).

2. E.g., models of censoring, truncation, etc.

3. Optimal theoretically, if specification is correct

4. Not robust (i.e., sensitive to distributional assumptions), a problem if model is incorrect

5. Often difficult practically
Application-specific Methods for Missing Data

1. Base inferences on the likelihood function or posterior distribution, by conditioning on observed data only, $P(\theta | Y_{obs})$.
2. E.g., models of censoring, truncation, etc.
3. Optimal theoretically, if specification is correct
4. Not robust (i.e., sensitive to distributional assumptions), a problem if model is incorrect
5. Often difficult practically
6. Very difficult with missingness scattered through $X$ and $Y$
How to create application-specific methods?

1. We observe $M$ always. Suppose we also see all the contents of $D$.

2. Then the likelihood is $P(D, M | \theta, \gamma) = P(D | \theta) P(M | D, \gamma)$, the likelihood if $D$ were observed, and the model for missingness.

3. Suppose now $D$ is observed (as usual) only when $M$ is 1.
How to create application-specific methods?

1. We observe $M$ always. Suppose we also see all the contents of $D$. 

   ▶ If $D$ and $M$ are observed, when can we drop $P(M | D, \gamma)$?

   ▶ Stochastic and parametric independence
How to create application-specific methods?

1. We observe $M$ always. Suppose we also see all the contents of $D$.
2. Then the likelihood is

$$P(D, M | \theta, \gamma) = P(D | \theta) P(M | D, \gamma),$$

the likelihood if $D$ were observed, and the model for missingness.

▶ If $D$ and $M$ are observed, when can we drop $P(M | D, \gamma)$?
1. We observe $M$ always. Suppose we also see all the contents of $D$.

2. Then the likelihood is

$$P(D, M|\theta, \gamma) = P(D|\theta)P(M|D, \gamma),$$
How to create application-specific methods?

1. We observe $M$ always. Suppose we also see all the contents of $D$.

2. Then the likelihood is

$$P(D, M|\theta, \gamma) = P(D|\theta)P(M|D, \gamma),$$

the likelihood if $D$ were observed, and the model for missingness.
How to create application-specific methods?

1. We observe $M$ always. Suppose we also see all the contents of $D$.

2. Then the likelihood is

$$P(D, M|\theta, \gamma) = P(D|\theta)P(M|D, \gamma),$$

the likelihood if $D$ were observed, and the model for missingness.

   ▶ If $D$ and $M$ are observed, when can we drop $P(M|D, \gamma)$?
How to create application-specific methods?

1. We observe $M$ always. Suppose we also see all the contents of $D$.
2. Then the likelihood is

$$P(D, M|\theta, \gamma) = P(D|\theta)P(M|D, \gamma),$$

the likelihood if $D$ were observed, and the model for missingness.

- If $D$ and $M$ are observed, when can we drop $P(M|D, \gamma)$?
- Stochastic and parametric independence
How to create application-specific methods?

1. We observe $M$ always. Suppose we also see all the contents of $D$.
2. Then the likelihood is

$$P(D, M|\theta, \gamma) = P(D|\theta)P(M|D, \gamma),$$

the likelihood if $D$ were observed, and the model for missingness.

- If $D$ and $M$ are observed, when can we drop $P(M|D, \gamma)$?
- Stochastic and parametric independence

3. Suppose now $D$ is observed (as usual) only when $M$ is 1.
4. Then the likelihood integrates out the missing observations

\[ P(D_{\text{obs}}, M | \theta, \gamma) = \int P(D | \theta) P(M | D, \gamma) \, dD_{\text{mis}} \]

and if we assume MAR (\( D \) and \( M \) are stochastically and parametrically independent), then

\[ P(D_{\text{obs}} | \theta) P(M | D_{\text{obs}}, \gamma) \propto P(D_{\text{obs}} | \theta) \]

because \( P(M | D_{\text{obs}}, \gamma) \) is constant w.r.t. \( \theta \).

5. Without the MAR assumption, the missingness model can't be dropped; it is NI (i.e., you can't ignore the model for \( M \)).

6. Specifying the missingness mechanism is hard. Little theory is available.

7. NI models (Heckman, many others) haven't always done well when the truth is known.
4. Then the likelihood integrates out the missing observations

\[ P(D_{obs}, M|\theta, \gamma) \]
4. Then the likelihood integrates out the missing observations

\[ P(D_{obs}, M|\theta, \gamma) = \int P(D|\theta)P(M|D, \gamma)dD_{mis} \]
4. Then the likelihood integrates out the missing observations

\[ P(D_{obs}, M|\theta, \gamma) = \int P(D|\theta)P(M|D, \gamma)dD_{mis} \]

and if assume MAR \((D \text{ and } M \text{ are stochastically and parametrically independent})\), then
4. Then the likelihood integrates out the missing observations

\[ P(D_{obs}, M|\theta, \gamma) = \int P(D|\theta)P(M|D, \gamma)dD_{mis} \]

and if assume MAR (\(D\) and \(M\) are stochastically and parametrically independent), then

\[ = \ P(D_{obs}|\theta)P(M|D_{obs}, \gamma), \]

5. Without the MAR assumption, the missingness model can't be dropped; it is NI (i.e., you can't ignore the model for \(M\)).

6. Specifying the missingness mechanism is hard. Little theory is available

7. NI models (Heckman, many others) haven't always done well when truth is known.
4. Then the likelihood integrates out the missing observations

\[ P(D_{obs}, M|\theta, \gamma) = \int P(D|\theta)P(M|D, \gamma)dD_{mis} \]

and if assume MAR (\( D \) and \( M \) are stochastically and parametrically independent), then

\[ = P(D_{obs}|\theta)P(M|D_{obs}, \gamma), \]
\[ \propto P(D_{obs}|\theta) \]
4. Then the likelihood integrates out the missing observations

\[ P(D_{obs}, M|\theta, \gamma) = \int P(D|\theta)P(M|D, \gamma)dD_{mis} \]

and if assume MAR \((D \text{ and } M \text{ are stochastically and parametrically independent})\), then

\[ = P(D_{obs}|\theta)P(M|D_{obs}, \gamma), \]

\[ \propto P(D_{obs}|\theta) \]

because \(P(M|D_{obs}, \gamma)\) is constant w.r.t. \(\theta\)
4. Then the likelihood integrates out the missing observations

\[
P(D_{obs}, M|\theta, \gamma) = \int P(D|\theta)P(M|D, \gamma)dD_{mis}
\]

and if assume MAR (\(D\) and \(M\) are stochastically and parametrically independent), then

\[
= P(D_{obs}|\theta)P(M|D_{obs}, \gamma), \\
\propto P(D_{obs}|\theta)
\]

because \(P(M|D_{obs}, \gamma)\) is constant w.r.t. \(\theta\)

5. Without the MAR assumption, the missingness model can’t be dropped; it is NI (i.e., you can’t ignore the model for \(M\))
4. Then the likelihood integrates out the missing observations

\[ P(D_{\text{obs}}, M|\theta, \gamma) = \int P(D|\theta)P(M|D, \gamma)dD_{\text{mis}} \]

and if assume MAR (\(D\) and \(M\) are stochastically and parametrically independent), then

\[ = P(D_{\text{obs}}|\theta)P(M|D_{\text{obs}}, \gamma), \]
\[ \propto P(D_{\text{obs}}|\theta) \]

because \(P(M|D_{\text{obs}}, \gamma)\) is constant w.r.t. \(\theta\)

5. Without the MAR assumption, the missingness model can’t be dropped; it is NI (i.e., you can’t ignore the model for \(M\))

6. Specifying the missingness mechanism is hard. Little theory is available
4. Then the likelihood integrates out the missing observations

\[ P(D_{obs}, M|\theta, \gamma) = \int P(D|\theta)P(M|D, \gamma)dD_{mis} \]

and if assume MAR (\(D\) and \(M\) are stochastically and parametrically independent), then

\[ = P(D_{obs}|\theta)P(M|D_{obs}, \gamma), \]

\[ \propto P(D_{obs}|\theta) \]

because \(P(M|D_{obs}, \gamma)\) is constant w.r.t. \(\theta\)

5. Without the MAR assumption, the missingness model can’t be dropped; it is NI (i.e., you can’t ignore the model for \(M\))

6. Specifying the missingness mechanism is hard. Little theory is available

7. NI models (Heckman, many others) haven’t always done well when truth is known
Strong General Purpose Missing Data Methods

- Maximum Likelihood
Strong General Purpose Missing Data Methods

- Maximum Likelihood
  - idea: maximize the observed log-likelihood
Strong General Purpose Missing Data Methods

- **Maximum Likelihood**
  - idea: maximize the observed log-likelihood
  - missingness model and analysis model are combined

- **Doubly Robust Weighting**
  - idea: combine weighting and regression imputation by reweighting on the residual
  - requires a model for the missingness propensity score and the regression imputation model
  - consistent if either of the above models is correct.

- **Multiple Imputation**
  - idea: compute several completed datasets
  - decouples analysis from the missingness model
  - what we will talk about primarily today
Strong General Purpose Missing Data Methods

- **Maximum Likelihood**
  - idea: maximize the observed log-likelihood
  - missingness model and analysis model are combined
  - great performance under MAR if data generating process is correct

- **Doubly Robust Weighting**
  - idea: combine weighting and regression imputation by reweighting on the residual
  - requires a model for the missingness propensity score and the regression imputation model
  - consistent if either of the above models is correct.
Strong General Purpose Missing Data Methods

- Maximum Likelihood
  - idea: maximize the observed log-likelihood
  - missingness model and analysis model are combined
  - great performance under MAR if data generating process is correct

- Doubly Robust Weighting
Strong General Purpose Missing Data Methods

- **Maximum Likelihood**
  - idea: maximize the observed log-likelihood
  - missingness model and analysis model are combined
  - great performance under MAR if data generating process is correct

- **Doubly Robust Weighting**
  - idea: combine weighting and regression imputation by reweighting on the residual
Strong General Purpose Missing Data Methods

- **Maximum Likelihood**
  - idea: maximize the observed log-likelihood
  - missingness model and analysis model are combined
  - great performance under MAR if data generating process is correct

- **Doubly Robust Weighting**
  - idea: combine weighting and regression imputation by reweighting on the residual
  - requires a model for the missingness propensity score and the regression imputation model
Strong General Purpose Missing Data Methods

- **Maximum Likelihood**
  - idea: maximize the observed log-likelihood
  - missingness model and analysis model are combined
  - great performance under MAR if data generating process is correct

- **Doubly Robust Weighting**
  - idea: combine weighting and regression imputation by reweighting on the residual
  - requires a model for the missingness propensity score and the regression imputation model
  - consistent if either of the above models is correct.
Strong General Purpose Missing Data Methods

- **Maximum Likelihood**
  - idea: maximize the observed log-likelihood
  - missingness model and analysis model are combined
  - great performance under MAR if data generating process is correct

- **Doubly Robust Weighting**
  - idea: combine weighting and regression imputation by reweighting on the residual
  - requires a model for the missingness propensity score and the regression imputation model
  - consistent if either of the above models is correct.

- **Multiple Imputation**
Strong General Purpose Missing Data Methods

- **Maximum Likelihood**
  - idea: maximize the observed log-likelihood
  - missingness model and analysis model are combined
  - great performance under MAR if data generating process is correct

- **Doubly Robust Weighting**
  - idea: combine weighting and regression imputation by reweighting on the residual
  - requires a model for the missingness propensity score and the regression imputation model
  - consistent if either of the above models is correct.

- **Multiple Imputation**
  - idea: compute several completed datasets
Strong General Purpose Missing Data Methods

- Maximum Likelihood
  - idea: maximize the observed log-likelihood
  - missingness model and analysis model are combined
  - great performance under MAR if data generating process is correct

- Doubly Robust Weighting
  - idea: combine weighting and regression imputation by reweighting on the residual
  - requires a model for the missingness propensity score and the regression imputation model
  - consistent if either of the above models is correct.

- Multiple Imputation
  - idea: compute several completed datasets
  - decouples analysis from the missingness model
Strong General Purpose Missing Data Methods

- **Maximum Likelihood**
  - idea: maximize the observed log-likelihood
  - missingness model and analysis model are combined
  - great performance under MAR if data generating process is correct

- **Doubly Robust Weighting**
  - idea: combine weighting and regression imputation by reweighting on the residual
  - requires a model for the missingness propensity score and the regression imputation model
  - consistent if either of the above models is correct.

- **Multiple Imputation**
  - idea: compute several completed datasets
  - decouples analysis from the missingness model
  - what we will talk about primarily today
1 Mixture Models
   - Basic Mixtures
   - Application: Mixtures as Preprocessing
   - Application: Mixture of Regressions

2 Expectation Maximization
   - EM for Probit Regression
   - EM for Gaussian Mixtures
   - EM in General

3 Missing Data
   - Motivating Example
   - Overview and Assumptions
   - Existing Heuristics
   - Application Specific Approaches
   - Multiple Imputation
   - The Full Amelia Scheme

4 Measurement Error

5 Appendix: Additional Details and Examples
1 Mixture Models
   • Basic Mixtures
   • Application: Mixtures as Preprocessing
   • Application: Mixture of Regressions

2 Expectation Maximization
   • EM for Probit Regression
   • EM for Gaussian Mixtures
   • EM in General

3 Missing Data
   • Motivating Example
   • Overview and Assumptions
   • Existing Heuristics
   • Application Specific Approaches
     • Multiple Imputation
   • The Full Amelia Scheme

4 Measurement Error

5 Appendix: Additional Details and Examples
Multiple Imputation

Point estimates are consistent, efficient, and the standard errors are right.

To compute:

1. Impute \( m \) values for each missing element
   - (a) Imputation method assumes MAR
   - (b) Uses a model with stochastic and systematic components
   - (c) Produces independent imputations
   - (d) (We'll give you a model to impute later)

2. Create \( m \) completed data sets
   - (a) Observed data are the same across the data sets
   - (b) Imputations of missing data differ
     - i. Cells we can predict well don't differ much
     - ii. Cells we can't predict well differ a lot

3. Run whatever statistical method you would have with no missing data for each completed data set
Multiple Imputation

Point estimates are consistent, efficient, and the standard errors are right. To compute:

1. Impute $m$ values for each missing element
   - (a) Imputation method assumes MAR
   - (b) Uses a model with stochastic and systematic components
   - (c) Produces independent imputations
   - (d) (We'll give you a model to impute later)

2. Create $m$ completed data sets
   - (a) Observed data are the same across the data sets
   - (b) Imputations of missing data differ
     - i. Cells we can predict well don't differ much
     - ii. Cells we can't predict well differ a lot

3. Run whatever statistical method you would have with no missing data for each completed data set
Multiple Imputation

Point estimates are consistent, efficient, and the standard errors are right. To compute:

1. Impute $m$ values for each missing element
Multiple Imputation

Point estimates are consistent, efficient, and the standard errors are right.
To compute:
1. Impute $m$ values for each missing element
   (a) Imputation method assumes MAR
Multiple Imputation

Point estimates are consistent, efficient, and the standard errors are right. To compute:

1. **Impute** $m$ values for each missing element
   - (a) Imputation method assumes MAR
   - (b) Uses a model with stochastic and systematic components
Multiple Imputation

Point estimates are consistent, efficient, and the standard errors are right.

To compute:

1. **Impute \( m \) values for each missing element**
   (a) Imputation method assumes MAR
   (b) Uses a model with stochastic and systematic components
   (c) Produces independent imputations
Multiple Imputation

Point estimates are consistent, efficient, and the standard errors are right. To compute:

1. **Impute $m$ values for each missing element**
   - (a) Imputation method assumes MAR
   - (b) Uses a model with stochastic and systematic components
   - (c) Produces independent imputations
   - (d) (We’ll give you a model to impute later)
Multiple Imputation

Point estimates are consistent, efficient, and the standard errors are right. To compute:

1. **Impute $m$ values for each missing element**
   (a) Imputation method assumes MAR
   (b) Uses a model with stochastic and systematic components
   (c) Produces independent imputations
   (d) (We’ll give you a model to impute later)

2. **Create $m$ completed data sets**
Multiple Imputation

Point estimates are consistent, efficient, and the standard errors are right. To compute:

1. **Impute $m$ values for each missing element**
   a. Imputation method assumes MAR
   b. Uses a model with stochastic and systematic components
   c. Produces independent imputations
   d. (We’ll give you a model to impute later)

2. **Create $m$ completed data sets**
   a. Observed data are the same across the data sets
Multiple Imputation

Point estimates are consistent, efficient, and the standard errors are right.

To compute:

1. **Impute $m$ values for each missing element**
   (a) Imputation method assumes MAR
   (b) Uses a model with stochastic and systematic components
   (c) Produces independent imputations
   (d) (We’ll give you a model to impute later)

2. **Create $m$ completed data sets**
   (a) Observed data are the same across the data sets
   (b) Imputations of missing data differ
Multiple Imputation

Point estimates are consistent, efficient, and the standard errors are right. To compute:

1. **Impute $m$ values for each missing element**
   (a) Imputation method assumes MAR
   (b) Uses a model with stochastic and systematic components
   (c) Produces independent imputations
   (d) (We’ll give you a model to impute later)

2. **Create $m$ completed data sets**
   (a) Observed data are the same across the data sets
   (b) Imputations of missing data differ
      i. Cells we can predict well don’t differ much
Multiple Imputation

Point estimates are consistent, efficient, and the standard errors are right. To compute:

1. **Impute* *$m$* values for each missing element**
   - (a) Imputation method assumes MAR
   - (b) Uses a model with stochastic and systematic components
   - (c) Produces independent imputations
   - (d) (We’ll give you a model to impute later)

2. **Create* *$m$* completed data sets**
   - (a) Observed data are the same across the data sets
   - (b) Imputations of missing data differ
     - i. Cells we can predict well don’t differ much
     - ii. Cells we can’t predict well differ a lot
Multiple Imputation

Point estimates are consistent, efficient, and the standard errors are right.

To compute:

1. **Impute $m$ values for each missing element**
   (a) Imputation method assumes MAR
   (b) Uses a model with stochastic and systematic components
   (c) Produces independent imputations
   (d) (We’ll give you a model to impute later)

2. **Create $m$ completed data sets**
   (a) Observed data are the same across the data sets
   (b) Imputations of missing data differ
      i. Cells we can predict well don’t differ much
      ii. Cells we can’t predict well differ a lot

3. **Run whatever statistical method you would have** with no missing data for each completed data set
4. Overall Point estimate: average individual point estimates, $q_j$

$(j = 1, \ldots, m)$:
4. **Overall Point estimate**: average individual point estimates, $q_j$

($j = 1, \ldots, m$):

$$ \bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j $$
4. Overall Point estimate: average individual point estimates, $q_j$
($j = 1, \ldots, m$):

$$\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j$$

Standard error: use Rubin’s Rule:
4. **Overall Point estimate**: average individual point estimates, \( q_j \) 
\((j = 1, \ldots, m)\):

\[
\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j
\]

**Standard error**: use Rubin’s Rule:

\[
SE(q)^2 = \text{mean}(SE_j^2) + \text{variance}(q_j) \left(1 + \frac{1}{m}\right)
\]
4. **Overall Point estimate**: average individual point estimates, \( q_j \) 
\((j = 1, \ldots, m)\):

\[
\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j
\]

**Standard error**: use Rubin’s Rule:

\[
SE(q)^2 = \text{mean}(SE_j^2) + \text{variance}(q_j)(1 + 1/m)
\]

\[
= \text{within + between}
\]
4. **Overall Point estimate**: average individual point estimates, $q_j$ 
($j = 1, \ldots, m$):

$$
\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j
$$

**Standard error**: use Rubin's Rule:

$$
SE(q)^2 = \text{mean}(SE_j^2) + \text{variance}(q_j) \left( 1 + \frac{1}{m} \right)
$$

$$
= \text{within} + \text{between}
$$

**Last piece vanishes as $m$ increases**
4. **Overall Point estimate**: average individual point estimates, $q_j$ ($j = 1, \ldots, m$):

$$\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j$$

**Standard error**: use Rubin’s Rule:

$$SE(q)^2 = \text{mean}(SE_j^2) + \text{variance}(q_j) \left(1 + \frac{1}{m}\right)$$

$$= \text{within} + \text{between}$$

*Last piece vanishes as $m$ increases*

5. **Easier by simulation**: draw $1/m$ sims from each data set of the QOI, combine (i.e., concatenate into a larger set of simulations), and make inferences as usual.
A General Model for Imputations

1. If data were complete, we could use (it's deceptively simple):

\[
L(\mu, \Sigma | D) \propto n \prod_{i=1}^{N} \mathcal{N}(D_i | \mu, \Sigma)
\]

(a multivariate normal)

2. With missing data, this becomes:

\[
L(\mu, \Sigma | D_{\text{obs}}) \propto n \prod_{i=1}^{N} \int \mathcal{N}(D_i | \mu, \Sigma) dD_{\text{mis}} = n \prod_{i=1}^{N} \mathcal{N}(D_i, \text{obs} | \mu_{\text{obs}}, \Sigma_{\text{obs}})
\]

since marginals of MVN's are normal.

3. Simple theoretically: merely a likelihood model for data \((D_{\text{obs}}, M)\) and same parameters as when fully observed \((\mu, \Sigma)\).

4. Difficult computationally: \(D_i, \text{obs}\) has different elements observed for each \(i\) and so each observation is informative about different pieces of \((\mu, \Sigma)\).
A General Model for Imputations

1. If data were complete, we could use (it's deceptively simple):
A General Model for Imputations

1. If data were complete, we could use (it's deceptively simple):

\[
L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma)
\]

2. With missing data, this becomes:

\[
L(\mu, \Sigma | D_{\text{obs}}) \propto \prod_{i=1}^{n} \int N(D_i | \mu, \Sigma) dD_{\text{mis}} = \prod_{i=1}^{n} N(D_i, \text{obs} | \mu_{\text{obs}}, \Sigma_{\text{obs}})
\]

since marginals of MVN's are normal.

3. Simple theoretically: merely a likelihood model for data (\(D_{\text{obs}}, M\)) and same parameters as when fully observed (\(\mu, \Sigma\)).

4. Difficult computationally: \(D_{\text{obs}}\) has different elements observed for each \(i\) and so each observation is informative about different pieces of (\(\mu, \Sigma\)).
A General Model for Imputations

1. If data were complete, we could use (it’s deceptively simple):

\[ L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \] (a multivariate normal)

2. With missing data, this becomes:

\[ L(\mu, \Sigma | D_{\text{obs}}) \propto \prod_{i=1}^{n} \int N(D_i | \mu, \Sigma) dD_{\text{mis}} = \prod_{i=1}^{n} N(D_i, \text{obs} | \mu_{\text{obs}}, \Sigma_{\text{obs}}) \] since marginals of MVN's are normal.

3. Simple theoretically: merely a likelihood model for data \((D_{\text{obs}}, M)\) and same parameters as when fully observed \((\mu, \Sigma)\).

4. Difficult computationally: \(D_{i, \text{obs}}\) has different elements observed for each \(i\) and so each observation is informative about different pieces of \((\mu, \Sigma)\).
A General Model for Imputations

1. If data were complete, we could use (it’s deceptively simple):

\[ L(\mu, \Sigma|D) \propto \prod_{i=1}^{n} N(D_i|\mu, \Sigma) \quad \text{ (a multivariate normal)} \]

2. With missing data, this becomes:

\[ L(\mu, \Sigma|D_{\text{obs}}) \propto \prod_{i=1}^{n} \int N(D_i|\mu, \Sigma) dD_{\text{mis}} = \prod_{i=1}^{n} N(D_i|\mu_{\text{obs}}, \Sigma_{\text{obs}}) \]

since marginals of MVN's are normal.
A General Model for Imputations

1. If data were complete, we could use (it’s deceptively simple):

\[ L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \]  
(a multivariate normal)

2. With missing data, this becomes:

\[ L(\mu, \Sigma | D_{obs}) \propto \prod_{i=1}^{n} \int N(D_i | \mu, \Sigma) dD_{mis} \]
A General Model for Imputations

1. If data were complete, we could use (it’s deceptively simple):

\[
L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \quad \text{(a multivariate normal)}
\]

2. With missing data, this becomes:

\[
L(\mu, \Sigma | D_{obs}) \propto \prod_{i=1}^{n} \int N(D_i | \mu, \Sigma) dD_{mis} = \prod_{i=1}^{n} N(D_{i,obs} | \mu_{obs}, \Sigma_{obs})
\]
A General Model for Imputations

1. If data were complete, we could use (it’s deceptively simple):

\[
L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \quad \text{(a multivariate normal)}
\]

2. With missing data, this becomes:

\[
L(\mu, \Sigma | D_{obs}) \propto \prod_{i=1}^{n} \int N(D_i | \mu, \Sigma) dD_{mis}
\]

\[
= \prod_{i=1}^{n} N(D_{i, obs} | \mu_{obs}, \Sigma_{obs})
\]

since marginals of MVN’s are normal.

Stewart (Princeton)
A General Model for Imputations

1. If data were complete, we could use (it’s deceptively simple):

\[
L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \quad (\text{a multivariate normal})
\]

2. With missing data, this becomes:

\[
L(\mu, \Sigma | D_{obs}) \propto \prod_{i=1}^{n} \int N(D_i | \mu, \Sigma) dD_{mis}
\]

\[
= \prod_{i=1}^{n} N(D_{i, obs} | \mu_{obs}, \Sigma_{obs})
\]

since marginals of MVN’s are normal.

3. Simple theoretically: merely a likelihood model for data \((D_{obs}, M)\) and same parameters as when fully observed \((\mu, \Sigma)\).
A General Model for Imputations

1. If data were complete, we could use (it’s deceptively simple):

\[ L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \]  

(a multivariate normal)

2. With missing data, this becomes:

\[ L(\mu, \Sigma | D_{obs}) \propto \prod_{i=1}^{n} \int N(D_i | \mu, \Sigma) dD_{mis} \]

\[ = \prod_{i=1}^{n} N(D_{i,obs} | \mu_{obs}, \Sigma_{obs}) \]

since marginals of MVN’s are normal.

3. Simple theoretically: merely a likelihood model for data \((D_{obs}, M)\) and same parameters as when fully observed \((\mu, \Sigma)\).

4. Difficult computationally: \(D_{i,obs}\) has different elements observed for each \(i\) and so each observation is informative about different pieces of \((\mu, \Sigma)\).
5. **Difficult Statistically**: number of parameters increases quickly in the number of variables \((p, \text{columns of } D)\):

\[
\text{parameters} = \text{parameters}(\mu) + \text{parameters}(\Sigma) = p + \frac{p(p+1)}{2} = \frac{p(p+3)}{2}.
\]

E.g., for \(p = 5\), parameters = 20; for \(p = 40\) parameters = 860 (Compare to \(n\)).

6. More appropriate models, such as for categorical or mixed variables, are harder to apply and do not usually perform better than this model (If you’re going to use a difficult imputation method, you might as well use an application-specific method. The goal is an easy-to-apply, generally applicable, method even if 2nd best.)

7. For social science survey data, which mostly contain ordinal scales, this is a reasonable choice for imputation, even though it may not be a good choice for analysis.
5. **Difficult Statistically**: number of parameters increases quickly in the number of variables ($p$, columns of $D$):

\[
\text{parameters} = \text{parameters}(\mu) + \text{parameters}(\Sigma)
\]
5. **Difficult Statistically**: number of parameters increases quickly in the number of variables ($p$, columns of $D$):

\[
\text{parameters} = \text{parameters}(\mu) + \text{parameters}(\Sigma) \\
= p + p(p + 1)/2
\]
5. **Difficult Statistically**: number of parameters increases quickly in the number of variables ($p$, columns of $D$):

\[
\text{parameters} = \text{parameters}(\mu) + \text{parameters}(\Sigma) \\
= p + p(p + 1)/2 = p(p + 3)/2.
\]
5. **Difficult Statistically**: number of parameters increases quickly in the number of variables ($p$, columns of $D$):

\[
\text{parameters} = \text{parameters}(\mu) + \text{parameters}(\Sigma) = p + p(p + 1)/2 = p(p + 3)/2.
\]

E.g., for $p = 5$, parameters$= 20$; for $p = 40$ parameters$= 860$ (Compare to $n$.)
5. **Difficult Statistically**: number of parameters increases quickly in the number of variables ($p$, columns of $D$):

\[
\text{parameters} = \text{parameters}(\mu) + \text{parameters}(\Sigma)
= p + p(p + 1)/2 = p(p + 3)/2.
\]

E.g., for $p = 5$, parameters = 20; for $p = 40$ parameters = 860 (Compare to $n$.)

6. **More appropriate models**, such as for categorical or mixed variables, are harder to apply and do not usually perform better than this model (If you’re going to use a difficult imputation method, you might as well use an application-specific method. The goal is an easy-to-apply, generally applicable, method even if 2nd best.)
5. **Difficult Statistically**: number of parameters increases quickly in the number of variables \((p, \text{columns of } D)\):

\[
\text{parameters} = \text{parameters}(\mu) + \text{parameters}(\Sigma) \\
= p + p(p + 1)/2 = p(p + 3)/2.
\]

E.g., for \(p = 5\), parameters\(= 20\); for \(p = 40\) parameters\(= 860\) (Compare to \(n\).)

6. **More appropriate models**, such as for categorical or mixed variables, are harder to apply and do not usually perform better than this model (If you’re going to use a difficult imputation method, you might as well use an application-specific method. The goal is an easy-to-apply, generally applicable, method even if 2nd best.)

7. **For social science survey data**, which mostly contain ordinal scales, this is a reasonable choice for imputation, even though it may not be a good choice for analysis.
How to create imputations from this model

1. E.g., suppose $D$ has only 2 variables, $D = \{X, Y\}$

2. $X$ is fully observed, $Y$ has some missingness.

3. Then $D = \{Y, X\}$ is bivariate normal:

$$D \sim N(D | \mu, \Sigma) = N([Y \ X] | \mu_Y \mu_X, \sigma_Y \sigma_X \sigma_{XY} \sigma_{XY} \sigma_X)$$

4. Conditionals of bivariate normals are normal:

$$Y | X \sim N(Y | E(Y | X), V(Y | X))$$

where

- $E(Y | X) = \mu_Y + \beta(X - \mu_X)$ (a regression of $Y$ on all other $X$'s!) $\beta = \sigma_{XY} / \sigma_X$
- $V(Y | X) = \sigma_Y ^2 - \sigma_{XY} ^2 / \sigma_X ^2$
How to create imputations from this model

1. E.g., suppose \( D \) has only 2 variables, \( D = \{X, Y\} \)
How to create imputations from this model

1. E.g., suppose $D$ has only 2 variables, $D = \{X, Y\}$
2. $X$ is fully observed, $Y$ has some missingness.
How to create imputations from this model

1. E.g., suppose $D$ has only 2 variables, $D = \{X, Y\}$
2. $X$ is fully observed, $Y$ has some missingness.
3. Then $D = \{Y, X\}$ is bivariate normal:

\begin{align*}
&D \sim N(D | \mu, \Sigma)
\end{align*}
How to create imputations from this model

1. E.g., suppose $D$ has only 2 variables, $D = \{X, Y\}$
2. $X$ is fully observed, $Y$ has some missingness.
3. Then $D = \{Y, X\}$ is bivariate normal:

$$D \sim N(D|\mu, \Sigma)$$
How to create imputations from this model

1. E.g., suppose $D$ has only 2 variables, $D = \{X, Y\}$
2. $X$ is fully observed, $Y$ has some missingness.
3. Then $D = \{Y, X\}$ is bivariate normal:

$$D \sim N(D | \mu, \Sigma)$$

$$= N \left[ \begin{pmatrix} Y \\ X \end{pmatrix} \middle| \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix} \right]$$
How to create imputations from this model

1. E.g., suppose $D$ has only 2 variables, $D = \{X, Y\}$
2. $X$ is fully observed, $Y$ has some missingness.
3. Then $D = \{Y, X\}$ is bivariate normal:
   \[
   D \sim N(D|\mu, \Sigma)
   \]
   \[
   = N\left[\begin{pmatrix} Y \\ X \end{pmatrix} \mid \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix}\right]
   \]
4. Conditionals of bivariate normals are normal:
How to create imputations from this model

1. E.g., suppose $D$ has only 2 variables, $D = \{X, Y\}$
2. $X$ is fully observed, $Y$ has some missingness.
3. Then $D = \{Y, X\}$ is bivariate normal:

$$D \sim N(D|\mu, \Sigma) = N \left[ \begin{pmatrix} Y \\ X \end{pmatrix} \middle| \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix} \right]$$

4. Conditionals of bivariate normals are normal:

$$Y|X \sim N(y|E(Y|X), V(Y|X))$$
How to create imputations from this model

1. E.g., suppose $D$ has only 2 variables, $D = \{X, Y\}$
2. $X$ is fully observed, $Y$ has some missingness.
3. Then $D = \{Y, X\}$ is bivariate normal:

$$D \sim N(D|\mu, \Sigma)$$

$$= N \left[ \begin{pmatrix} Y \\ X \end{pmatrix}, \begin{pmatrix} \mu_Y \\ \mu_X, \sigma_y \\ \sigma_x \end{pmatrix} \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix} \right]$$

4. Conditionals of bivariate normals are normal:

$$Y|X \sim N(y|E(Y|X), V(Y|X))$$

$$E(Y|X) = \mu_y + \beta(X - \mu_x) \quad \text{(a regression of } Y \text{ on all other } X\text{'s!})$$
How to create imputations from this model

1. E.g., suppose $D$ has only 2 variables, $D = \{X, Y\}$
2. $X$ is fully observed, $Y$ has some missingness.
3. Then $D = \{Y, X\}$ is bivariate normal:

   \[
   D \sim N(D|\mu, \Sigma)
   \]

   \[
   = N \left[ \begin{pmatrix} Y \\ X \end{pmatrix} \bigg| \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix} \right]
   \]

4. Conditionals of bivariate normals are normal:

   $Y|X \sim N(y|E(Y|X), V(Y|X))$

   \[
   E(Y|X) = \mu_y + \beta(X - \mu_x) \quad \text{(a regression of $Y$ on all other $X$'s!)}
   \]

   \[
   \beta = \frac{\sigma_{xy}}{\sigma_x}
   \]
How to create imputations from this model

1. E.g., suppose $D$ has only 2 variables, $D = \{X, Y\}$
2. $X$ is fully observed, $Y$ has some missingness.
3. Then $D = \{Y, X\}$ is bivariate normal:

$$D \sim N(D|\mu, \Sigma)$$

$$= N \begin{bmatrix} (Y) \\ (X) \end{bmatrix} \bar{|} \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix}$$

4. Conditionals of bivariate normals are normal:

$$Y|X \sim N(y|E(Y|X), V(Y|X))$$

$$E(Y|X) = \mu_y + \beta(X - \mu_x) \quad \text{(a regression of } Y \text{ on all other } X\text{'s!)}$$

$$\beta = \sigma_{xy}/\sigma_x$$

$$V(Y|X) = \sigma_y - \sigma_{xy}^2/\sigma_x$$
5. To create imputations:

(a) Estimate the posterior density of $\mu$ and $\Sigma$. Could do the usual: maximize likelihood, assume CLT applies, and draw from the normal approximation. (Hard to do, and CLT isn't a good asymptotic approximation due to large number of parameters.)

(b) Draw $\mu$ and $\Sigma$ from their posterior density

(c) Compute simulations of $E(Y|X)$ and $V(Y|X)$ deterministically

(d) Draw a simulation of the missing $Y$ from the conditional normal

6. In this simple example ($X$ fully observed), this is equivalent to simulating from a linear regression of $Y$ on $X$, $\tilde{y}_i = x_i \tilde{\beta} + \tilde{\epsilon}_i$, with estimation and fundamental uncertainty
5. **To create imputations:**
   
   (a) Estimate the posterior density of $\mu$ and $\Sigma$

   ii. We will improve on this shortly

   (b) Draw $\mu$ and $\Sigma$ from their posterior density

   (c) Compute simulations of $E(Y|X)$ and $V(Y|X)$ deterministically

   (d) Draw a simulation of the missing $Y$ from the conditional normal
5. To create imputations:
(a) Estimate the posterior density of $\mu$ and $\Sigma$
   i. Could do the usual: maximize likelihood, assume CLT applies, and draw from the normal approximation. (Hard to do, and CLT isn’t a good asymptotic approximation due to large number of parameters.)
5. **To create imputations:**

   (a) Estimate the posterior density of $\mu$ and $\Sigma$

      i. Could do the usual: maximize likelihood, assume CLT applies, and draw from the normal approximation. (Hard to do, and CLT isn’t a good asymptotic approximation due to large number of parameters.)

      ii. We will improve on this shortly

---

6. In this simple example ($X$ fully observed), this is equivalent to simulating from a linear regression of $Y$ on $X$, $\tilde{y}_i = x_i \tilde{\beta} + \tilde{\epsilon}_i$, with estimation and fundamental uncertainty
5. To create imputations:
   (a) Estimate the posterior density of $\mu$ and $\Sigma$
      i. Could do the usual: maximize likelihood, assume CLT applies, and draw from the normal approximation. (Hard to do, and CLT isn’t a good asymptotic approximation due to large number of parameters.)
      ii. We will improve on this shortly
   (b) Draw $\mu$ and $\Sigma$ from their posterior density

6. In this simple example ($X$ fully observed), this is equivalent to simulating from a linear regression of $Y$ on $X$, 
   $\tilde{y}_i = x_i \tilde{\beta} + \tilde{\epsilon}_i$,
5. **To create imputations:**
   
   (a) Estimate the posterior density of $\mu$ and $\Sigma$
      
      i. Could do the usual: maximize likelihood, assume CLT applies, and draw from the normal approximation. (Hard to do, and CLT isn’t a good asymptotic approximation due to large number of parameters.)
      
      ii. We will improve on this shortly
   
   (b) Draw $\mu$ and $\Sigma$ from their posterior density
   
   (c) Compute simulations of $E(Y|X)$ and $V(Y|X)$ deterministically

6. In this simple example ($X$ fully observed), this is equivalent to simulating from a linear regression of $Y$ on $X$, 
   
   $\tilde{y}_i = x_i \tilde{\beta} + \tilde{\epsilon}_i$, 

   with estimation and fundamental uncertainty.
5. **To create imputations:**

(a) Estimate the posterior density of $\mu$ and $\Sigma$
   
   i. Could do the usual: maximize likelihood, assume CLT applies, and draw from the normal approximation. (Hard to do, and CLT isn’t a good asymptotic approximation due to large number of parameters.)
   
   ii. We will improve on this shortly

(b) Draw $\mu$ and $\Sigma$ from their posterior density

(c) Compute simulations of $E(Y|X)$ and $V(Y|X)$ deterministically

(d) Draw a simulation of the missing $Y$ from the conditional normal
5. To create imputations:
   (a) Estimate the posterior density of $\mu$ and $\Sigma$
      i. Could do the usual: maximize likelihood, assume CLT applies, and draw from the normal approximation. (Hard to do, and CLT isn’t a good asymptotic approximation due to large number of parameters.)
      ii. We will improve on this shortly
   (b) Draw $\mu$ and $\Sigma$ from their posterior density
   (c) Compute simulations of $E(Y|X)$ and $V(Y|X)$ deterministically
   (d) Draw a simulation of the missing $Y$ from the conditional normal

6. In this simple example ($X$ fully observed), this is equivalent to simulating from a linear regression of $Y$ on $X$,

$$\tilde{y}_i = x_i \tilde{\beta} + \tilde{\epsilon}_i,$$

with estimation and fundamental uncertainty
Randomly draw $n$ obs (with replacement) from the data
Use EM to estimate $\beta$ and $\Sigma$ in each (for estimation uncertainty)
Impute $D_{mis}$ from each (for fundamental uncertainty)

Lightning fast; works with very large data sets

Basis for Amelia II

Stewart (Princeton)
Randomly draw $n$ obs (with replacement) from the data
EMB: EM With Bootstrap

- Randomly draw $n$ obs (with replacement) from the data
- Use EM to estimate $\beta$ and $\Sigma$ in each (for estimation uncertainty)
Randomly draw $n$ obs (with replacement) from the data

Use EM to estimate $\beta$ and $\Sigma$ in each (for estimation uncertainty)

Impute $D_{mis}$ from each from the model (for fundamental uncertainty)
EMB: EM With Bootstrap

- Randomly draw $n$ obs (with replacement) from the data
- Use EM to estimate $\beta$ and $\Sigma$ in each (for estimation uncertainty)
- Impute $D_{mis}$ from each from the model (for fundamental uncertainty)
- Lightning fast; works with very large data sets
Randomly draw $n$ obs (with replacement) from the data
Use EM to estimate $\beta$ and $\Sigma$ in each (for estimation uncertainty)
Impute $D_{mis}$ from each from the model (for fundamental uncertainty)
Lightning fast; works with very large data sets
Basis for Amelia II
Multiple Imputation: Amelia Style
Multiple Imputation: Amelia Style

incomplete data
Multiple Imputation: Amelia Style

- incomplete data
- bootstrap
- bootstrapped data
Multiple Imputation: Amelia Style

- Incomplete data
- Bootstrap
- Bootstrapped data
- EM
- Imputed datasets

Stewart (Princeton)

Missing Data

Mar 27-Apr 5, 2017 79 / 200
Multiple Imputation: Amelia Style

- Incomplete data
- Bootstrap
- Bootstrapped data
- EM
- Imputed datasets
- Analysis
Multiple Imputation: Amelia Style

- incomplete data
- bootstrap
- bootstrapped data
- EM
- imputed datasets
- analysis
- combination
- final results
What Can Go Wrong and What to Do

Inference is learning about facts we don’t have with facts we have; we assume the 2 are related!

Imputation and analysis are estimated separately; robustness because imputation affects only missing observations. High missingness reduces the property.

Include at least as much information in the imputation model as in the analysis model: all vars in analysis model; others that would help predict (e.g., All measures of a variable, post-treatment variables).

Fit imputation model distributional assumptions by transformation to unbounded scales: \( \sqrt{\text{counts}} \), \( \ln\left(\frac{p}{1-p}\right) \), \( \ln(\text{money}) \), etc.

Code ordinal variables as close to interval as possible.
What Can Go Wrong and What to Do

- Inference is learning about facts we don’t have with facts we have; we assume the 2 are related!

- Imputation and analysis are estimated separately; robustness because imputation affects only missing observations. High missingness reduces the property.

- Include at least as much information in the imputation model as in the analysis model: all vars in analysis model; others that would help predict (e.g., All measures of a variable, post-treatment variables)

- Fit imputation model distributional assumptions by transformation to unbounded scales: \( \sqrt{\text{counts}} \), \( \ln(p/(1-p)) \), \( \ln(\text{money}) \), etc.

- Code ordinal variables as close to interval as possible.
What Can Go Wrong and What to Do

- Inference is learning about facts we don’t have with facts we have; we assume the 2 are related!
- Imputation and analysis are estimated separately $\rightsquigarrow$ robustness because imputation affects only missing observations. High missingness reduces the property.

- Include at least as much information in the imputation model as in the analysis model: all vars in analysis model; others that would help predict (e.g., All measures of a variable, post-treatment variables)
- Fit imputation model distributional assumptions by transformation to unbounded scales: $\sqrt{\text{counts}}, \ln\left(\frac{p}{1-p}\right), \ln(\text{money}),$ etc.
- Code ordinal variables as close to interval as possible.
What Can Go Wrong and What to Do

- Inference is learning about facts we don’t have with facts we have; we assume the 2 are related!
- Imputation and analysis are estimated separately \( \sim \) robustness because imputation affects only missing observations. High missingness reduces the property.
- Include at least as much information in the imputation model as in the analysis model: all vars in analysis model; others that would help predict (e.g., All measures of a variable, post-treatment variables)

Stewart (Princeton)
What Can Go Wrong and What to Do

- Inference is learning about facts we don’t have with facts we have; we assume the 2 are related!
- Imputation and analysis are estimated separately \( \rightsquigarrow \) robustness because imputation affects only missing observations. High missingness reduces the property.
- Include at least as much information in the imputation model as in the analysis model: all vars in analysis model; others that would help predict (e.g., All measures of a variable, post-treatment variables)
- Fit imputation model distributional assumptions by transformation to unbounded scales: \( \sqrt{\text{counts}}, \ln(\frac{p}{1-p}), \ln(\text{money}), \text{etc.} \)
What Can Go Wrong and What to Do

- Inference is learning about facts we don’t have with facts we have; we assume the 2 are related!
- Imputation and analysis are estimated separately → robustness because imputation affects only missing observations. High missingness reduces the property.
- Include at least as much information in the imputation model as in the analysis model: all vars in analysis model; others that would help predict (e.g., All measures of a variable, post-treatment variables)
- Fit imputation model distributional assumptions by transformation to unbounded scales: \( \sqrt{\text{counts}} \), \( \ln(p/(1-p)) \), \( \ln(\text{money}) \), etc.
- Code ordinal variables as close to interval as possible.
What Can Go Wrong and What to Do (continued)

Represent severe nonlinear relationships in the imputation model with transformations or added quadratic terms.

If imputation model has as much information as the analysis model, but the specification (such as the functional form) differs, CIs are conservative (e.g., \(\geq 95\% \text{ CIs}\)).

When imputation model includes more information than analysis model, it can be more efficient than the “optimal” application-specific model (known as “super-efficiency”).

Bad intuitions

- If \(X\) is randomly imputed why no attenuation (the usual consequence of random measurement error in an explanatory variable)?
- If \(X\) is imputed with information from \(Y\), why no endogeneity?
- Answer to both: the draws are from the joint posterior and put back into the data. Nothing is being changed.
Represent severe nonlinear relationships in the imputation model with transformations or added quadratic terms.
What Can Go Wrong and What to Do (continued)

- Represent severe nonlinear relationships in the imputation model with transformations or added quadratic terms.
- If imputation model has as much information as the analysis model, but the specification (such as the functional form) differs, CIs are conservative (e.g., $\geq 95\%$ CIs)

Additional points:

- If $X$ is randomly imputed why no attenuation (the usual consequence of random measurement error in an explanatory variable)?
- If $X$ is imputed with information from $Y$, why no endogeneity?
- Answer to both: the draws are from the joint posterior and put back into the data. Nothing is being changed.
What Can Go Wrong and What to Do (continued)

- Represent severe nonlinear relationships in the imputation model with transformations or added quadratic terms.
- If imputation model has as much information as the analysis model, but the specification (such as the functional form) differs, CIs are conservative (e.g., $\geq 95\%$ CIs)
- When imputation model includes more information than analysis model, it can be more efficient than the “optimal” application-specific model (known as “super-efficiency”)

Bad intuitions
- If $X$ is randomly imputed why no attenuation (the usual consequence of random measurement error in an explanatory variable)?
- If $X$ is imputed with information from $Y$, why no endogeneity?
- Answer to both: the draws are from the joint posterior and put back into the data. Nothing is being changed.
What Can Go Wrong and What to Do (continued)

- Represent severe nonlinear relationships in the imputation model with transformations or added quadratic terms.
- If imputation model has as much information as the analysis model, but the specification (such as the functional form) differs, CIs are conservative (e.g., ≥ 95% CIs)
- When imputation model includes more information than analysis model, it can be more efficient than the “optimal” application-specific model (known as “super-efficiency”)
- Bad intuitions

▶ If $X$ is randomly imputed why no attenuation (the usual consequence of random measurement error in an explanatory variable)?
▶ If $X$ is imputed with information from $Y$, why no endogeneity?

Answer to both: the draws are from the joint posterior and put back into the data. Nothing is being changed.
What Can Go Wrong and What to Do (continued)

- Represent severe nonlinear relationships in the imputation model with transformations or added quadratic terms.
- If imputation model has as much information as the analysis model, but the specification (such as the functional form) differs, CIs are conservative (e.g., ≥ 95% CIs)
- When imputation model includes more information than analysis model, it can be more efficient than the “optimal” application-specific model (known as “super-efficiency”)
- Bad intuitions
  - If $X$ is randomly imputed why no attenuation (the usual consequence of random measurement error in an explanatory variable)?
What Can Go Wrong and What to Do (continued)

- Represent severe nonlinear relationships in the imputation model with transformations or added quadratic terms.
- If imputation model has as much information as the analysis model, but the specification (such as the functional form) differs, CIs are conservative (e.g., ≥ 95% CIs)
- When imputation model includes more information than analysis model, it can be more efficient than the “optimal” application-specific model (known as “super-efficiency”)
- Bad intuitions
  - If $X$ is randomly imputed why no attenuation (the usual consequence of random measurement error in an explanatory variable)?
  - If $X$ is imputed with information from $Y$, why no endogeneity?
What Can Go Wrong and What to Do (continued)

- Represent severe nonlinear relationships in the imputation model with transformations or added quadratic terms.
- If imputation model has as much information as the analysis model, but the specification (such as the functional form) differs, CIs are conservative (e.g., $\geq 95\%$ CIs)
- When imputation model includes more information than analysis model, it can be more efficient than the “optimal” application-specific model (known as “super-efficiency”)
- Bad intuitions
  - If $X$ is randomly imputed why no attenuation (the usual consequence of random measurement error in an explanatory variable)?
  - If $X$ is imputed with information from $Y$, why no endogeneity?
  - Answer to both: the draws are from the joint posterior and put back into the data. Nothing is being changed.
1 Mixture Models
   - Basic Mixtures
   - Application: Mixtures as Preprocessing
   - Application: Mixture of Regressions

2 Expectation Maximization
   - EM for Probit Regression
   - EM for Gaussian Mixtures
   - EM in General

3 Missing Data
   - Motivating Example
   - Overview and Assumptions
   - Existing Heuristics
   - Application Specific Approaches
   - Multiple Imputation
   - The Full Amelia Scheme

4 Measurement Error

5 Appendix: Additional Details and Examples
1 Mixture Models
   - Basic Mixtures
   - Application: Mixtures as Preprocessing
   - Application: Mixture of Regressions

2 Expectation Maximization
   - EM for Probit Regression
   - EM for Gaussian Mixtures
   - EM in General

3 Missing Data
   - Motivating Example
   - Overview and Assumptions
   - Existing Heuristics
   - Application Specific Approaches
   - Multiple Imputation
   - The Full Amelia Scheme

4 Measurement Error

5 Appendix: Additional Details and Examples
What Does AMELIA II Do?

The AMELIA II algorithm begins by assuming that the functional form of the complete data is multivariate normal:

\[(y, X) \sim MVN(\mu, \Sigma).\]
What Does AMELIA II Do?

The AMELIA II algorithm begins by assuming that the functional form of the complete data is multivariate normal:

$$(y, X) \sim MVN(\mu, \Sigma).$$

Let’s define $(y,X) = D$ to simplify the notation.
What Does AMELIA II Do?

The AMELIA II algorithm begins by assuming that the functional form of the complete data is multivariate normal:

$$(y, X) \sim \text{MVN}(\mu, \Sigma).$$

Let’s define $(y, X) = D$ to simplify the notation.

Once again this gives us two full conditionals for our unknowns $(\mu, \Sigma, D_{mis})$:

1. $p(D_{Mis} | \mu, \Sigma, D_{Obs})$
2. $L(\mu, \Sigma | D_{Obs}, D_{Mis})$
What Does AMELIA II Do?

The AMELIA II algorithm begins by assuming that the functional form of the complete data is multivariate normal:

\[(y, X) \sim MVN(\mu, \Sigma).\]

Let’s define \((y, X) = D\) to simplify the notation.

Once again this gives us two full conditionals for our unknowns \((\mu, \Sigma, D_{mis})\):

1. \(p(D_{Mis}|\mu, \Sigma, D_{Obs})\)
2. \(L(\mu, \Sigma|D_{Obs}, D_{Mis})\)

The EM algorithm in this case involves selecting an initial value for \((\mu, \Sigma)\), using that value to impute the missing data, and then re-estimating \((\mu, \Sigma)\) based on the (now-complete) data.
The Multiple Imputation Scheme
The Multiple Imputation Scheme

incomplete data
The Multiple Imputation Scheme

incomplete data
imputation
imputed datasets
The Multiple Imputation Scheme

- **incomplete data**
- **imputation**
- **imputed datasets**
- **analysis**
- **separate results**
The Multiple Imputation Scheme

- Incomplete data
- Imputation
- Imputed datasets
- Analysis
- Separate results
- Combination
- Final results
Multiple Imputation

To preserve the relationships in the data.

To reflect the uncertainty of our imputation.
Multiple Imputation

REGRESSION
To preserve the relationships in the data.
Multiple Imputation

REGRESSION
To preserve the relationships in the data.

SIMULATION
To reflect the uncertainty of our imputation.
How to Impute

\[ y = X\hat{\beta} + \varepsilon \]
How to Impute

\[ X_i^{\text{mis}} = X_i^{\text{obs}} \hat{\beta} + \hat{\epsilon} \]

REGRESSION
**How to Impute**

\[
X_i^{\text{mis}} = X_i^{\text{obs}} \hat{\beta} + \hat{\epsilon}
\]

**REGRESSION**

\[
\hat{\beta} \sim \mathcal{N}(\beta, \text{var}(\hat{\beta})) \quad \text{SIMULATION} \quad \hat{\epsilon} \sim \mathcal{N}(0, \hat{\sigma}_X^2)
\]
## Patterns of Missingness

<table>
<thead>
<tr>
<th>year</th>
<th>country</th>
<th>GDP</th>
<th>infl</th>
<th>trade</th>
<th>population</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Burkina Faso</td>
<td>377</td>
<td>-2.92</td>
<td>29.69</td>
<td>5848380</td>
</tr>
<tr>
<td>2</td>
<td>Burkina Faso</td>
<td>376</td>
<td>7.60</td>
<td>31.31</td>
<td>5958700</td>
</tr>
<tr>
<td>3</td>
<td>Burkina Faso</td>
<td>393</td>
<td>NA</td>
<td>NA</td>
<td>6075700</td>
</tr>
<tr>
<td>4</td>
<td>Burkina Faso</td>
<td>416</td>
<td>18.76</td>
<td>40.11</td>
<td>6202000</td>
</tr>
<tr>
<td>5</td>
<td>Burkina Faso</td>
<td>435</td>
<td>NA</td>
<td>37.76</td>
<td>6341030</td>
</tr>
<tr>
<td>6</td>
<td>Burkina Faso</td>
<td>448</td>
<td>29.99</td>
<td>41.11</td>
<td>6486870</td>
</tr>
</tbody>
</table>

\[ infl = \beta_0 + \beta_1 \cdot GDP + \beta_2 \cdot population + \varepsilon \]
## Patterns of Missingness

<table>
<thead>
<tr>
<th>year</th>
<th>country</th>
<th>GDP</th>
<th>infl</th>
<th>trade</th>
<th>population</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Burkina Faso</td>
<td>377</td>
<td>-2.92</td>
<td>29.69</td>
<td>5848380</td>
</tr>
<tr>
<td>2</td>
<td>Burkina Faso</td>
<td>376</td>
<td>7.60</td>
<td>31.31</td>
<td>5958700</td>
</tr>
<tr>
<td>3</td>
<td>Burkina Faso</td>
<td>393</td>
<td>NA</td>
<td>NA</td>
<td>6075700</td>
</tr>
<tr>
<td>4</td>
<td>Burkina Faso</td>
<td>416</td>
<td>18.76</td>
<td>40.11</td>
<td>6202000</td>
</tr>
<tr>
<td>5</td>
<td>Burkina Faso</td>
<td>435</td>
<td>NA</td>
<td>37.76</td>
<td>6341030</td>
</tr>
<tr>
<td>6</td>
<td>Burkina Faso</td>
<td>448</td>
<td>29.99</td>
<td>41.11</td>
<td>6486870</td>
</tr>
</tbody>
</table>

\[
\text{trade} = \beta_0 + \beta_1 \cdot GDP + \beta_2 \cdot \text{population} + \varepsilon
\]
Patterns of Missingness

<table>
<thead>
<tr>
<th>year</th>
<th>country</th>
<th>GDP</th>
<th>infl</th>
<th>trade</th>
<th>population</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1972 Burkina Faso</td>
<td>377</td>
<td>-2.92</td>
<td>29.69</td>
<td>5848380</td>
</tr>
<tr>
<td>2</td>
<td>1973 Burkina Faso</td>
<td>376</td>
<td>7.60</td>
<td>31.31</td>
<td>5958700</td>
</tr>
<tr>
<td>3</td>
<td>1974 Burkina Faso</td>
<td>393</td>
<td>NA</td>
<td>NA</td>
<td>6075700</td>
</tr>
<tr>
<td>4</td>
<td>1975 Burkina Faso</td>
<td>416</td>
<td>18.76</td>
<td>40.11</td>
<td>6202000</td>
</tr>
<tr>
<td>5</td>
<td>1976 Burkina Faso</td>
<td>435</td>
<td>NA</td>
<td>37.76</td>
<td>6341030</td>
</tr>
<tr>
<td>6</td>
<td>1977 Burkina Faso</td>
<td>448</td>
<td>29.99</td>
<td>41.11</td>
<td>6486870</td>
</tr>
</tbody>
</table>

\[
\text{infl} = \beta_0 + \beta_1 \cdot \text{GDP} + \beta_2 \cdot \text{trade} + \beta_3 \cdot \text{population} + \varepsilon
\]
Any $\beta$ is just $(\mu, \Sigma)$

- If $X \sim \mathcal{N}(\mu, \Sigma)$, we can recover any regression from the vector of means and the covariance matrix.
- Thus, we need $(\mu, \Sigma|X^{\text{obs}})$. 
A complicated likelihood

\[ \mathcal{L}(\mu, \Sigma | D^{\text{obs}}) \propto \prod_{i=1}^{n} \mathcal{N}(D_i^{\text{obs}} | \mu_i^{\text{obs}}, \Sigma_i^{\text{obs}}) \]
The EM algorithm

Turn a hard problem into a repeated easy problem.

1. Use current estimates of \((\mu, \Sigma)\) to estimate \(X^{\text{mis}}\).
2. Use those estimates of \(X^{\text{mis}}\) and \(X^{\text{obs}}\) to get a new estimate of \((\mu, \Sigma)\).
3. Iterate until convergence.

\[(\mu_t, \Sigma_t)\]
The EM algorithm

Turn a hard problem into a repeated easy problem.

1. Use current estimates of \((\mu, \Sigma)\) to estimate \(X^{mis}\).
2. Use those estimates of \(X^{mis}\) and \(X^{obs}\) to get a new estimate of \((\mu, \Sigma)\).
3. Iterate until convergence.

\[
E[X_{t+1}^{mis}]
\]

\[(\mu_t, \Sigma_t)\]
The EM algorithm

Turn a hard problem into a repeated easy problem.

1. Use current estimates of \((\mu, \Sigma)\) to estimate \(X^{\text{mis}}\).
2. Use those estimates of \(X^{\text{mis}}\) and \(X^{\text{obs}}\) to get a new estimate of \((\mu, \Sigma)\).
3. Iterate until convergence.

\[
\mathbb{E}[X_{t+1}^{\text{mis}}] \\
(\mu_t, \Sigma_t) \\
(\mu_{t+1}, \Sigma_{t+1}) \\
X^{\text{obs}}
\]
The EM algorithm

Turn a hard problem into a repeated easy problem.

1. Use current estimates of \((\mu, \Sigma)\) to estimate \(X^{\text{mis}}\).
2. Use those estimates of \(X^{\text{mis}}\) and \(X^{\text{obs}}\) to get a new estimate of \((\mu, \Sigma)\).
3. Iterate until convergence.

\[
\begin{align*}
\mathbb{E}[X_{t+1}^{\text{mis}}] & \quad \mathbb{E}[X_{t+2}^{\text{mis}}] \\
(\mu_t, \Sigma_t) & \quad (\mu_{t+1}, \Sigma_{t+1}) \\
& \quad X^{\text{obs}}
\end{align*}
\]
The EM algorithm

Turn a hard problem into a repeated easy problem.

1. Use current estimates of \((\mu, \Sigma)\) to estimate \(X^{\text{mis}}\).
2. Use those estimates of \(X^{\text{mis}}\) and \(X^{\text{obs}}\) to get a new estimate of \((\mu, \Sigma)\).
3. Iterate until convergence.

\[
(\mu_t, \Sigma_t) \quad \xrightarrow{\mathbb{E}[X^{\text{mis}}]} \quad (\mu_{t+1}, \Sigma_{t+1}) \quad \xrightarrow{\mathbb{E}[X^{\text{mis}}]} \quad (\mu_{t+2}, \Sigma_{t+2})
\]

\[
X^{\text{obs}} \quad \xrightarrow{\mathbb{E}[X^{\text{mis}}]} \quad X^{\text{obs}}
\]
Simulation

\[ X_i^{\text{mis}} = X_i^{\text{obs}} \beta + \varepsilon \]

missing values in row i
observed values in row i

EM is a tool for **REGRESSION**. In order to **SIMULATE**, we need...

1. a Normal approximation.
2. importance sampling.
3. a bootstrap-based approach.
How to Impute

\[ X_{i}^{\text{mis}} = X_{i}^{\text{obs}} \hat{\beta} + \hat{\varepsilon} \]

EM

\[ \hat{\beta} \sim \mathcal{N}(\beta, \text{var}(\hat{\beta})) \]

SIMULATION

\[ \hat{\varepsilon} \sim \mathcal{N}(0, \hat{\sigma}^2_{X_{\text{mis}}}) \]
How to Impute

\[ X_{i}^{\text{mis}} = X_{i}^{\text{obs}} \hat{\beta} + \hat{\epsilon} \]

EM

\[ \hat{\beta} \sim \mathcal{N}(\beta, \text{var}(\hat{\beta})) \]

BOOTSTRAP

\[ \hat{\epsilon} \sim \mathcal{N}(0, \hat{\sigma}^{2}_{X_{\text{mis}}}) \]
How to Impute

\[ X_i^{\text{mis}} = X_i^{\text{obs}} \beta + \varepsilon \]

where

- \( X_i^{\text{mis}} \) represents the missing values in row \( i \)
- \( X_i^{\text{obs}} \) represents the observed values in row \( i \)
- \( \varepsilon \sim \mathcal{N}(0, \sigma^2) \)

\( \beta \) is the parameter to be estimated.
How to Impute

$X_i^{\text{mis}} = X_i^{\text{obs}} \beta + \varepsilon \sim \mathcal{N}(0, \sigma^2)$

where $X_i^{\text{mis}}$ represents missing values in row $i$, $X_i^{\text{obs}}$ represents observed values in row $i$, $\beta$ are the regression coefficients, and $\varepsilon$ is the error term with a normal distribution $\mathcal{N}(0, \sigma^2)$. This is the regression imputation method for handling missing data.
How to Impute

\[ X_i^{\text{mis}} = X_i^{\text{obs}} \beta + \varepsilon \]

- \( X_i^{\text{mis}} \): missing values in row i
- \( X_i^{\text{obs}} \): observed values in row i
- \( \beta \): regression coefficients
- \( \varepsilon \): simulation error

\( \varepsilon \sim \mathcal{N}(0, \sigma^2) \)
How to Impute

- We will impute a missing value by drawing from a Normal distribution centered around what it's predicted by a regression of that variable on the available data in that observation.
How to Impute

- We will impute a missing value by drawing from a Normal distribution centered around what its predicted by a regression of that variable on the available data in that observation.
- A hard part is the regression, as we have to run a regression for every missing value in every pattern of missingness.
How to Impute

- We will impute a missing value by drawing from a Normal distribution centered around what its predicted by a regression of that variable on the available data in that observation.
- A hard part is the regression, as we have to run a regression for every missing value in every pattern of missingness.
- This could be a lot of regressions, depending on the data.
The Amelia Scheme
The Amelia Scheme

incomplete data
The Amelia Scheme

- **Incomplete data**
- **Bootstrap**
- **Bootstrapped data**
The Amelia Scheme

- **incomplete data**
- **bootstrap**
- **bootstrapped data**
- **EM**
- **imputed datasets**
The Amelia Scheme

- incomplete data
- bootstrap
- bootstrapped data
- EM
- imputed datasets
- analysis
The Amelia Scheme

- incomplete data
- bootstrap
- bootstrapped data
- EM
- imputed datasets
- analysis
- combination
- final results
The Amelia approach

1. Draw a sample of size $n$ with replacement, $X^*$. 
The Amelia approach

1. Draw a sample of size $n$ with replacement, $X^*$.  
2. Run the EM algorithm on $X^*$ the bootstrapped data to get estimates $(\hat{\mu}^*, \hat{\Sigma}^*)$. 
The *Amelia* approach

1. Draw a sample of size $n$ with replacement, $X^*$.
2. Run the EM algorithm on $X^*$ the bootstrapped data to get estimates $(\hat{\mu}^*, \hat{\Sigma}^*)$.
3. Use $(\hat{\mu}^*, \hat{\Sigma}^*)$ to impute the original data, $X$. 
The *Amelia* approach

1. Draw a sample of size $n$ with replacement, $X^*$.
2. Run the EM algorithm on $X^*$ the bootstrapped data to get estimates ($\hat{\mu}^*, \hat{\Sigma}^*$).
3. Use ($\hat{\mu}^*, \hat{\Sigma}^*$) to impute the original data, $X$.
4. Iterate $m$ times.
Assumptions and Estimators for Missing Data

- It is important to distinguish the estimator from the assumptions necessary to identify the model.
Assumptions and Estimators for Missing Data

- It is important to distinguish the estimator from the assumptions necessary to identify the model.
- Assumptions are typically some form of ignorability (MCAR, MAR, NMAR) and cannot be directly checked.

Recent work on diagnostics for multiple imputation provides us a place to start.

Not really covered here but see the Amelia vignette and the Su et al. paper.
Assumptions and Estimators for Missing Data

- It is important to distinguish the estimator from the assumptions necessary to identify the model.
- Assumptions are typically some form of ignorability (MCAR, MAR, NMAR) and cannot be directly checked.
- It is often more feasible to check for errors in estimation where we typically assume:
  - Sample is large enough for ML estimate to be approximately unbiased and normally distributed.
  - Parametric likelihood model is a decent approximation to the complete likelihood.

Recent work on diagnostics for multiple imputation provides us a place to start.
Not really covered here but see the Amelia vignette and the Su et al. paper.
Assumptions and Estimators for Missing Data

- It is important to distinguish the estimator from the assumptions necessary to identify the model.
- Assumptions are typically some form of ignorability (MCAR, MAR, NMAR) and cannot be directly checked.
- It is often more feasible to check for errors in estimation where we typically assume:
  - Sample is large enough for ML estimate to be approximately unbiased and normally distributed.
Assumptions and Estimators for Missing Data

- It is important to distinguish the estimator from the assumptions necessary to identify the model.
- Assumptions are typically some form of ignorability (MCAR, MAR, NMAR) and cannot be directly checked.
- It is often more feasible to check for errors in estimation where we typically assume:
  - sample is large enough for ML estimate to be approximately unbiased and normally distributed.
  - parametric likelihood model is a decent approximation to the complete likelihood.

Recent work on diagnostics for multiple imputation provides us a place to start.
Not really covered here but see the Amelia vignette and the Su et al paper.
Assumptions and Estimators for Missing Data

- It is important to distinguish the estimator from the assumptions necessary to identify the model.
- Assumptions are typically some form of ignorability (MCAR, MAR, NMAR) and cannot be directly checked.
- It is often more feasible to check for errors in estimation where we typically assume:
  - Sample is large enough for ML estimate to be approximately unbiased and normally distributed.
  - Parametric likelihood model is a decent approximation to the complete likelihood.
- Recent work on diagnostics for multiple imputation provides us a place to start.
Assumptions and Estimators for Missing Data

- It is important to distinguish the estimator from the assumptions necessary to identify the model.
- Assumptions are typically some form of ignorability (MCAR, MAR, NMAR) and cannot be directly checked.
- It is often more feasible to check for errors in estimation where we typically assume:
  - sample is large enough for ML estimate to be approximately unbiased and normally distributed
  - parametric likelihood model is a decent approximation to the complete likelihood
- Recent work on diagnostics for multiple imputation provides us a place to start.
- Not really covered here but see the Amelia vignette and the Su et al paper.
Example Amelia Diagnostics

Missingness Map

- Burkina Faso
- Burundi
- Cameroon
- Congo
- Senegal
- Zambia

- trade
- gdp_pc
- population
- civilib
- infl
- country
- year

Stewart (Princeton)
Missing Data
Mar 27-Apr 5, 2017 100 / 200
Example Amelia Diagnostics

Observed and Imputed values of gdp_pc

Relative Density

0.000 0.015

0 500 1500 2500

Observed and Imputed values of trade

gdp_pc   −−  Fraction Missing: 0.017

Observed and Imputed values of trade

Stewart (Princeton)
Example Amelia Diagnostics

Cameroon

Trade by Time (1975-1990)

Missing Data

Stewart (Princeton)

Mar 27-Apr 5, 2017 100 / 200
Example Amelia Diagnostics

Observed versus Imputed Values of trad

Stewart (Princeton)

Mar 27-Apr 5, 2017  100 / 200
Final Thoughts on Missing Data

- There is a bit of a goldilocks region here - too few observations and it doesn’t matter, too many and it will give crazy answers
Final Thoughts on Missing Data

- There is a bit of a goldilocks region here- too few observations and it doesn’t matter, too many and it will give crazy answers
- Overimputing and observed vs. imputed distributions are helpful diagnostics but there are no hard and fast rules
Final Thoughts on Missing Data

- There is a bit of a goldilocks region here- too few observations and it doesn’t matter, too many and it will give crazy answers.
- Overimputing and observed vs. imputed distributions are helpful diagnostics but there are no hard and fast rules.
- As per usual, domain knowledge here is key. The missing data literature just helps you apply that domain knowledge.
1 Mixture Models
   - Basic Mixtures
   - Application: Mixtures as Preprocessing
   - Application: Mixture of Regressions

2 Expectation Maximization
   - EM for Probit Regression
   - EM for Gaussian Mixtures
   - EM in General

3 Missing Data
   - Motivating Example
   - Overview and Assumptions
   - Existing Heuristics
   - Application Specific Approaches
   - Multiple Imputation
   - The Full Amelia Scheme

4 Measurement Error

5 Appendix: Additional Details and Examples
1 Mixture Models
   - Basic Mixtures
   - Application: Mixtures as Preprocessing
   - Application: Mixture of Regressions

2 Expectation Maximization
   - EM for Probit Regression
   - EM for Gaussian Mixtures
   - EM in General

3 Missing Data
   - Motivating Example
   - Overview and Assumptions
   - Existing Heuristics
   - Application Specific Approaches
   - Multiple Imputation
   - The Full Amelia Scheme

4 Measurement Error

5 Appendix: Additional Details and Examples
Three New Things

1. Measurement error is deeply problematic for political science research and current approaches are incorrect or unused.

2. Missing data is the limiting, most extreme form of measurement error.

3. We can rework the multiple imputation framework to simultaneously correct for both missing data and measurement error.
Three New Things

1. Measurement error is deeply problematic for political science research and current approaches are incorrect or unused.
Three New Things

1. Measurement error is deeply problematic for political science research and current approaches are incorrect or unused.
2. Missing data is the limiting, most extreme form of measurement error.
Three New Things

1. Measurement error is deeply problematic for political science research and current approaches are incorrect or unused.
2. Missing data is the limiting, most extreme form of measurement error.
3. We can rework the multiple imputation framework to simultaneously correct for both missing data and measurement error.
<table>
<thead>
<tr>
<th>country</th>
<th>polityiv</th>
<th>f-house</th>
<th>log-gdppc</th>
<th>primary</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUKINA FASO</td>
<td>5</td>
<td>6</td>
<td>6.23</td>
<td>5.92</td>
</tr>
<tr>
<td>LIBERIA</td>
<td>NA</td>
<td>3</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>SIERRA LEONE</td>
<td>3</td>
<td>3</td>
<td>6.60</td>
<td>NA</td>
</tr>
<tr>
<td>GHANA</td>
<td>9</td>
<td>6</td>
<td>6.86</td>
<td>12.68</td>
</tr>
<tr>
<td>TOGO</td>
<td>NA</td>
<td>5</td>
<td>6.27</td>
<td>17.34</td>
</tr>
<tr>
<td>CAMEROON</td>
<td>6</td>
<td>5</td>
<td>6.93</td>
<td>15.47</td>
</tr>
<tr>
<td>NIGERIA</td>
<td>5</td>
<td>7</td>
<td>6.88</td>
<td>17.46</td>
</tr>
<tr>
<td>GABON</td>
<td>6</td>
<td>8</td>
<td>8.19</td>
<td>16.97</td>
</tr>
<tr>
<td>country</td>
<td>polityiv</td>
<td>f-house</td>
<td>log-gdppc</td>
<td>primary</td>
</tr>
<tr>
<td>------------</td>
<td>----------</td>
<td>---------</td>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>BUKINA FASO</td>
<td>5</td>
<td>6</td>
<td>6.23</td>
<td>5.92</td>
</tr>
<tr>
<td>LIBERIA</td>
<td>NA</td>
<td>3</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>SIERRA LEONE</td>
<td>3</td>
<td>3</td>
<td>6.60</td>
<td>NA</td>
</tr>
<tr>
<td>GHANA</td>
<td>9</td>
<td>6</td>
<td>6.86</td>
<td>12.68</td>
</tr>
<tr>
<td>TOGO</td>
<td>NA</td>
<td>5</td>
<td>6.27</td>
<td>17.34</td>
</tr>
<tr>
<td>CAMEROON</td>
<td>6</td>
<td>5</td>
<td>6.93</td>
<td>15.47</td>
</tr>
<tr>
<td>NIGERIA</td>
<td>≈5</td>
<td>7</td>
<td>6.88</td>
<td>17.46</td>
</tr>
<tr>
<td>GABON</td>
<td>6</td>
<td>8</td>
<td>8.19</td>
<td>16.97</td>
</tr>
<tr>
<td>country</td>
<td>polityiv</td>
<td>f-house</td>
<td>log-gdppc</td>
<td>primary</td>
</tr>
<tr>
<td>-----------------</td>
<td>----------</td>
<td>---------</td>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>BUKINA FASO</td>
<td>≈5</td>
<td>6</td>
<td>6.23</td>
<td>5.92</td>
</tr>
<tr>
<td>LIBERIA</td>
<td>NA</td>
<td>3</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>SIERRA LEONE</td>
<td>≈3</td>
<td>3</td>
<td>6.60</td>
<td>NA</td>
</tr>
<tr>
<td>GHANA</td>
<td>≈9</td>
<td>6</td>
<td>6.86</td>
<td>12.68</td>
</tr>
<tr>
<td>TOGO</td>
<td>NA</td>
<td>5</td>
<td>6.27</td>
<td>17.34</td>
</tr>
<tr>
<td>CAMEROON</td>
<td>≈6</td>
<td>5</td>
<td>6.93</td>
<td>15.47</td>
</tr>
<tr>
<td>NIGERIA</td>
<td>≈5</td>
<td>7</td>
<td>6.88</td>
<td>17.46</td>
</tr>
<tr>
<td>GABON</td>
<td>≈6</td>
<td>8</td>
<td>8.19</td>
<td>16.97</td>
</tr>
<tr>
<td>country</td>
<td>polityiv</td>
<td>f-house</td>
<td>log-gdppc</td>
<td>primary</td>
</tr>
<tr>
<td>------------------</td>
<td>----------</td>
<td>---------</td>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>BUKINA FASO</td>
<td>≈5</td>
<td>≈6</td>
<td>6.23</td>
<td>5.92</td>
</tr>
<tr>
<td>LIBERIA</td>
<td>NA</td>
<td>≈3</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>SIERRA LEONE</td>
<td>≈3</td>
<td>≈3</td>
<td>6.60</td>
<td>NA</td>
</tr>
<tr>
<td>GHANA</td>
<td>≈9</td>
<td>≈6</td>
<td>6.86</td>
<td>12.68</td>
</tr>
<tr>
<td>TOGO</td>
<td>NA</td>
<td>≈6</td>
<td>6.27</td>
<td>17.34</td>
</tr>
<tr>
<td>CAMEROON</td>
<td>≈6</td>
<td>≈5</td>
<td>6.93</td>
<td>15.47</td>
</tr>
<tr>
<td>NIGERIA</td>
<td>≈5</td>
<td>≈7</td>
<td>6.88</td>
<td>17.46</td>
</tr>
<tr>
<td>GABON</td>
<td>≈6</td>
<td>≈8</td>
<td>8.19</td>
<td>16.97</td>
</tr>
</tbody>
</table>
One Solution:

<table>
<thead>
<tr>
<th>country</th>
<th>GDP</th>
<th>infl</th>
<th>trade</th>
<th>population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ghana</td>
<td>377</td>
<td>-2.92</td>
<td>29.69</td>
<td>5848380</td>
</tr>
<tr>
<td>Ivory Coast</td>
<td>376</td>
<td>7.60</td>
<td>31.31</td>
<td>5958700</td>
</tr>
<tr>
<td>Kenya</td>
<td>393</td>
<td>8.72</td>
<td>35.22</td>
<td>6075700</td>
</tr>
<tr>
<td>Nigeria</td>
<td>416</td>
<td>18.76</td>
<td>40.11</td>
<td>6202000</td>
</tr>
<tr>
<td>Uganda</td>
<td>435</td>
<td>-8.40</td>
<td>37.76</td>
<td>6341030</td>
</tr>
<tr>
<td>Burkina Faso</td>
<td>448</td>
<td>29.99</td>
<td>41.11</td>
<td>6486870</td>
</tr>
</tbody>
</table>
One Solution: Change research agendas

<table>
<thead>
<tr>
<th>country</th>
<th>GDP</th>
<th>infl</th>
<th>trade</th>
<th>population</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Ghana</td>
<td>377</td>
<td>-2.92</td>
<td>29.69</td>
<td>5848380</td>
</tr>
<tr>
<td>2 Ivory Coast</td>
<td>376</td>
<td>7.60</td>
<td>31.31</td>
<td>5958700</td>
</tr>
<tr>
<td>3 Kenya</td>
<td>393</td>
<td>8.72</td>
<td>35.22</td>
<td>6075700</td>
</tr>
<tr>
<td>4 Nigeria</td>
<td>416</td>
<td>18.76</td>
<td>40.11</td>
<td>6202000</td>
</tr>
<tr>
<td>5 Uganda</td>
<td>435</td>
<td>-8.40</td>
<td>37.76</td>
<td>6341030</td>
</tr>
<tr>
<td>6 Burkina Faso</td>
<td>448</td>
<td>29.99</td>
<td>41.11</td>
<td>6486870</td>
</tr>
</tbody>
</table>
One Solution: Change research agendas

<table>
<thead>
<tr>
<th>country</th>
<th>infl</th>
<th>trade</th>
<th>population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ghana</td>
<td>-2.92</td>
<td>29.69</td>
<td>5848380</td>
</tr>
<tr>
<td>Ivory Coast</td>
<td>7.60</td>
<td>31.31</td>
<td>5958700</td>
</tr>
<tr>
<td>Kenya</td>
<td>8.72</td>
<td>35.22</td>
<td>6075700</td>
</tr>
<tr>
<td>Nigeria</td>
<td>18.76</td>
<td>40.11</td>
<td>6202000</td>
</tr>
<tr>
<td>Uganda</td>
<td>-8.40</td>
<td>37.76</td>
<td>6341030</td>
</tr>
<tr>
<td>Burkina Faso</td>
<td>29.99</td>
<td>41.11</td>
<td>6486870</td>
</tr>
</tbody>
</table>
The current approaches in the literature

- Instrumental variables
The current approaches in the literature

- Instrumental variables
- Regression calibration
The current approaches in the literature

- Instrumental variables
- Regression calibration
- SIMEX
The current approaches in the literature

- Instrumental variables
- Regression calibration
- SIMEX
- Semiparametric models
The current approaches in the literature

- Instrumental variables
- Regression calibration
- SIMEX
- Semiparametric models
- Mixture models
The current approaches in the literature

- Instrumental variables
- Regression calibration
- SIMEX
- Semiparametric models
- Mixture models
- Quasi-likelihood models
The current approaches in the literature

- Instrumental variables
- Regression calibration
- SIMEX
- Semiparametric models
- Mixture models
- Quasi-likelihood models
- Denial
The current approaches in the literature

Most existing approaches are

application-specific.
model dependent.
difficult to implement.
inapplicable with multiple variables.
invalid with heteroskedastic errors.
unusable with missing data.
The current approaches in the literature

Most existing approaches are application-specific.
The current approaches in the literature

Most existing approaches are application-specific. model dependent.
The current approaches in the literature

Most existing approaches are application-specific.
model dependent.
difficult to implement.
The current approaches in the literature

Most existing approaches are application-specific. model dependent. difficult to implement. inapplicable with multiple variables.
The current approaches in the literature

Most existing approaches are application-specific. model dependent. difficult to implement. inapplicable with multiple variables. invalid with heteroskedastic errors.
The current approaches in the literature

Most existing approaches are application-specific. model dependent. difficult to implement. inapplicable with multiple variables. invalid with heteroskedastic errors. unusable with missing data.
Why is this the state of the art?
Why is this the state of the art?
It's easy and tolerated.
Why is this the state of the art?
It’s easy and tolerated.
But it’s make believe.
A Brief Review of Measurement Error

\[ x_i = x_i^* + u_i \]
A Brief Review of Measurement Error

\[ x_i = x_i^* + u_i \]
A Brief Review of Measurement Error

\[ x_i = x_i^* + u_i \]

- **observed**
- **latent**
A Brief Review of Measurement Error

\[ x_i = x_i^* + u_i \]

- \( x_i \): observed
- \( x_i^* \): latent
- \( u_i \): measurement error
A Brief Review of Measurement Error

\[ x_i = x_i^* + u_i \]

\[ u_i | x_i^* \sim \mathcal{N}(0, \sigma_u^2) \]
A Brief Review of Measurement Error

\[ x_i = x_i^* + u_i \]

\[ u_i | x_i^* \sim \mathcal{N}(0, \sigma_u^2) \]

observed \hspace{1cm} latent \hspace{1cm} measurement error

unbiased \hspace{1cm} independent
A Brief Review of Measurement Error

\[ x_i = x_i^* + u_i \]

\[ u_i | x_i^* \sim \mathcal{N}(0, \sigma_u^2) \]

observed \hspace{2cm} latent \hspace{2cm} measurement error

unbiased independent measurement error variance
Want to run:

\[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i \]

Can only run:

\[ y_i = \alpha_0 + \alpha_1 x_i + \nu_i \]

Leads to:

ATTENUATION (No guarantees with more mismeasured variables)

Stewart (Princeton)
Want to run:

\[ y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i \]
Want to run:

\[ y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i \]

Can only run:

\[ y_i = \alpha_0 + \alpha_1 x_i + \nu_i \]

Leads to: ATTENUATION (No guarantees with more mismeasured variables)
Want to run:

\[ y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i \]

Can only run:

\[ y_i = \alpha_0 + \alpha_1 x_i + \nu_i \]
Want to run:

\[ y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i \]

Can only run:

\[ y_i = \alpha_0 + \alpha_1 x_i + \nu_i \]

Leads to:
Want to run:

\[ y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i \]

Can only run:

\[ y_i = \alpha_0 + \alpha_1 x_i + \nu_i \]

Leads to:

ATTENUATION
Want to run:

\[ y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i \]

Can only run:

\[ y_i = \alpha_0 + \alpha_1 x_i + \nu_i \]

Leads to:

ATTENUATION

(But ONLY in linear models with one bad variable)
Want to run:

\[ y_i = \beta_0 + \beta_1 x_i^* + \beta_2 w_i^* + \beta_3 z_i^* + \epsilon_i \]

Can only run:

\[ y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 w_i + \alpha_3 z_i + \nu_i \]

Leads to:

**UNKNOWN**

*(No guarantees with more mismeasured variables)*
\[ \sigma_u^2 = 0 \]
$\sigma_u^2 = 0.1$
\[ \sigma_u^2 = 0.2 \]
\[ \sigma_u^2 = 0.4 \]
$\sigma^2_u = 0.6$
\[ \sigma_u^2 = 0.8 \]
ATTENUATION

...only guaranteed in the simplest of cases:
ATTENUATION

...only guaranteed in the simplest of cases:

linear model
ATTENUATION

...only guaranteed in the simplest of cases:

- linear model
- one mismeasured variable
ATTENUATION
...only guaranteed in the simplest of cases:

- linear model
- one mismeasured variable
- measurement error unrelated to other variables and $x^*$. 
BIAS FROM MEASUREMENT ERROR

In unpredictable directions with most realistic models.
The strict dichotomy of data.
The false dichotomy of data.

observed

missing
(fully) observed  (fully) missing
(fully) observed     (fully) missing

The false dichotomy of data.
Stewart (Princeton)

Missing Data

Mar 27-Apr 5, 2017 126 / 200
observed

fully observed

fully missing

Mar 27-Apr 5, 2017 126 / 200
Stewart (Princeton)

Missing Data

Mar 27-Apr 5, 2017

126 / 200
But what is this continuum?
\[ x_i = x_i^* + u_i \]

\[ u_i | x_i^* \sim \mathcal{N}(0, \sigma_u^2) \]
Fully observed \rightarrow \text{partially missing} \rightarrow \text{fully missing}
\[ \sigma_u^2 = 0 \]
\[
x_i = x_i^* + u_i
\]

\[
u_i \sim \mathcal{N}(0,0)
\]
\[ x_i = x_i^* + 0 \]

\[ u_i \sim \mathcal{N}(0,0) \]
$\sigma_u^2 = 0$

Diagram:
- Fully observed
- Partially missing
- Fully missing

Stewart (Princeton) Mar 27-Apr 5, 2017
$$\sigma^2_u = 0$$

$$\sigma^2_u = \infty$$
$$\sigma_u^2 = 0$$

$$0 < \sigma_u^2 < \infty$$

$$\sigma_u^2 = \infty$$
Missing data is the most extreme case of measurement error.
$X_i^*$
\[ x_i^* \]

- **fully observed**

\[ \sigma_u^2 \]

- **fully missing**
\[ x_i \]

\[ \sigma^2_u \]
NA

Fully observed

\( \sigma_u^2 \)

Fully missing
Multiple imputation:

observed  missing
Multiple imputation:

(fully) observed  (fully) missing
Multiple overimputation:

\[ x_i^* \quad \text{fully observed} \quad | \quad \text{fully missing} \]
Multiple overimputation:

\[
\begin{array}{ccc}
    x_i^* & \text{fully observed} & \text{partially missing} & \text{fully missing} \\
\end{array}
\]
Multiple overimputation:

\( x_i^* \quad \text{fully observed} \quad \text{perfectly measured} \quad \text{partially missing} \quad \text{measured with error} \quad \text{fully missing} \quad \text{infinite error} \)
Multiple overimputation:

\[
\begin{align*}
&x_i^* & \text{fully observed} & N(x_i^*, 0) \\
p(x_i|x_i^*) & \text{perfectly measured} & \text{partially missing} & \mathcal{N}(x_i^*, \sigma_u^2) \\
& & \text{measured with error} & \mathcal{N}(x_i^*, \infty) \\
& & \text{fully missing} \\
\end{align*}
\]
Multiple Overimputation extends the multiple imputation framework to correct for measurement error.
APPLICATION-SPECIFIC METHODS:
Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

incomplete
mismeasured
dataset

HARD!
EASY!
ROBUST!

Stewart (Princeton)
APPLICATION-SPECIFIC METHODS:

incomplete mismeasured dataset

missing data + measurement error + analysis

HARD!

EASY!

ROBUST!
Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

incomplete
mismeasured
dataset

missing data +
measurement error +
analysis

results
APPLICATION-SPECIFIC METHODS:

incomplete data + mismeasured data + analysis → results

HARD!
Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

incomplete mismeasured dataset

missing data + measurement error + analysis

MULTIPLE OVERIMPUTATION:

HARD!
Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

incomplete mismeasured dataset → missing data + measurement error + analysis → results

MULTIPLE OVERIMPUTATION:

incomplete mismeasured dataset

HARD!
Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

incomplete mismeasured dataset → missing data + measurement error + analysis → results

MULTIPLE OVERIMPUTATION:

incomplete mismeasured dataset → missing data + measurement error
Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

incomplete mismeasured dataset → missing data + measurement error + analysis → results

MULTIPLE OVERIMPUTATION:

incomplete mismeasured dataset → missing data + measurement error → analysis
Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

incomplete mismeasured dataset → missing data + measurement error + analysis → results

MULTIPLE OVERIMPUTATION:

incomplete mismeasured dataset → missing data + measurement error → analysis → results
APPLICATION-SPECIFIC METHODS:

incomplete mismeasured dataset → missing data + measurement error + analysis → results

MULTIPLE OVERIMPUTATION:

incomplete mismeasured dataset → missing data + measurement error → analysis → results

HARD!

EASY!
Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

incomplete mismeasured dataset → missing data + measurement error + analysis → results

HARD!

MULTIPLE OVERIMPUTATION:

incomplete mismeasured dataset → missing data + measurement error → analysis → results

ROBUST!

EASY!
What MO allows you to do:
What MO allows you to do:

social science.
<table>
<thead>
<tr>
<th>country</th>
<th>polityiv</th>
<th>f-house</th>
<th>log-gdppc</th>
<th>primary</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUKINA FASO</td>
<td>≈9</td>
<td>6</td>
<td>6.23</td>
<td>5.92</td>
</tr>
<tr>
<td>LIBERIA</td>
<td>NA</td>
<td>3</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>SIERRA LEONE</td>
<td>≈3</td>
<td>3</td>
<td>6.60</td>
<td>NA</td>
</tr>
<tr>
<td>GHANA</td>
<td>≈9</td>
<td>6</td>
<td>6.86</td>
<td>12.68</td>
</tr>
<tr>
<td>TOGO</td>
<td>NA</td>
<td>5</td>
<td>6.27</td>
<td>17.34</td>
</tr>
<tr>
<td>CAMEROON</td>
<td>≈6</td>
<td>5</td>
<td>6.93</td>
<td>15.47</td>
</tr>
<tr>
<td>NIGERIA</td>
<td>≈5</td>
<td>7</td>
<td>6.88</td>
<td>17.46</td>
</tr>
<tr>
<td>GABON</td>
<td>≈6</td>
<td>8</td>
<td>8.19</td>
<td>16.97</td>
</tr>
<tr>
<td>country</td>
<td>polityiv</td>
<td>f-house</td>
<td>log-gdppc</td>
<td>primary</td>
</tr>
<tr>
<td>--------------</td>
<td>----------</td>
<td>---------</td>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>BUKINA FASO</td>
<td>9</td>
<td>6</td>
<td>6.23</td>
<td>5.92</td>
</tr>
<tr>
<td>LIBERIA</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIERRA LEONE</td>
<td>3</td>
<td>3</td>
<td>6.60</td>
<td></td>
</tr>
<tr>
<td>GHANA</td>
<td>9</td>
<td>6</td>
<td>6.86</td>
<td>12.68</td>
</tr>
<tr>
<td>TOGO</td>
<td></td>
<td>5</td>
<td>6.27</td>
<td>17.34</td>
</tr>
<tr>
<td>CAMEROON</td>
<td>6</td>
<td>5</td>
<td>6.93</td>
<td>15.47</td>
</tr>
<tr>
<td>NIGERIA</td>
<td>5</td>
<td>7</td>
<td>6.88</td>
<td>17.46</td>
</tr>
<tr>
<td>GABON</td>
<td>6</td>
<td>8</td>
<td>8.19</td>
<td>16.97</td>
</tr>
<tr>
<td>country</td>
<td>polityiv</td>
<td>f-house</td>
<td>log-gdppc</td>
<td>primary</td>
</tr>
<tr>
<td>-----------------</td>
<td>----------</td>
<td>---------</td>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>BUKINA FASO</td>
<td></td>
<td>6</td>
<td>6.23</td>
<td>5.92</td>
</tr>
<tr>
<td>LIBERIA</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIERRA LEONE</td>
<td></td>
<td>3</td>
<td>6.60</td>
<td></td>
</tr>
<tr>
<td>GHANA</td>
<td></td>
<td>6</td>
<td>6.86</td>
<td>12.68</td>
</tr>
<tr>
<td>TOGO</td>
<td></td>
<td>5</td>
<td>6.27</td>
<td>17.34</td>
</tr>
<tr>
<td>CAMEROON</td>
<td></td>
<td>5</td>
<td>6.93</td>
<td>15.47</td>
</tr>
<tr>
<td>NIGERIA</td>
<td></td>
<td>7</td>
<td>6.88</td>
<td>17.46</td>
</tr>
<tr>
<td>GABON</td>
<td></td>
<td>8</td>
<td>8.19</td>
<td>16.97</td>
</tr>
<tr>
<td>country</td>
<td>polityiv</td>
<td>f-house</td>
<td>log-gdppc</td>
<td>primary</td>
</tr>
<tr>
<td>---------------</td>
<td>----------</td>
<td>---------</td>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>BUKINA FASO</td>
<td></td>
<td>6</td>
<td>6.23</td>
<td>5.92</td>
</tr>
<tr>
<td>LIBERIA</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIERRA LEONE</td>
<td></td>
<td>3</td>
<td>6.60</td>
<td></td>
</tr>
<tr>
<td>GHANA</td>
<td></td>
<td>6</td>
<td>6.86</td>
<td>12.68</td>
</tr>
<tr>
<td>TOGO</td>
<td></td>
<td>5</td>
<td>6.27</td>
<td>17.34</td>
</tr>
<tr>
<td>CAMEROON</td>
<td></td>
<td>5</td>
<td>6.93</td>
<td>15.47</td>
</tr>
<tr>
<td>NIGERIA</td>
<td></td>
<td>7</td>
<td>6.88</td>
<td>17.46</td>
</tr>
<tr>
<td>GABON</td>
<td></td>
<td>8</td>
<td>8.19</td>
<td>16.97</td>
</tr>
</tbody>
</table>

Stewart (Princeton)  
Mar 27-Apr 5, 2017  
142 / 200
<table>
<thead>
<tr>
<th>country</th>
<th>polityiv</th>
<th>f-house</th>
<th>log-gdppc</th>
<th>primary</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUKINA FASO</td>
<td>6</td>
<td>6.23</td>
<td>5.92</td>
<td></td>
</tr>
<tr>
<td>LIBERIA</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIERRA LEONE</td>
<td>3</td>
<td>6.60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GHANA</td>
<td>6</td>
<td>6.86</td>
<td>12.68</td>
<td></td>
</tr>
<tr>
<td>TOGO</td>
<td>5</td>
<td>6.27</td>
<td>17.34</td>
<td></td>
</tr>
<tr>
<td>CAMEROON</td>
<td>5</td>
<td>6.93</td>
<td>15.47</td>
<td></td>
</tr>
<tr>
<td>NIGERIA</td>
<td>7</td>
<td>6.88</td>
<td>17.46</td>
<td></td>
</tr>
<tr>
<td>GABON</td>
<td>8</td>
<td>8.19</td>
<td>16.97</td>
<td></td>
</tr>
<tr>
<td>country</td>
<td>polityiv</td>
<td>f-house</td>
<td>log-gdppc</td>
<td>primary</td>
</tr>
<tr>
<td>---------------</td>
<td>----------</td>
<td>---------</td>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>BUKINA FASO</td>
<td>▼</td>
<td>6</td>
<td>6.23</td>
<td>5.92</td>
</tr>
<tr>
<td>LIBERIA</td>
<td>▼</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIERRA LEONE</td>
<td>▼</td>
<td>3</td>
<td>6.60</td>
<td></td>
</tr>
<tr>
<td>GHANA</td>
<td>▼</td>
<td>6</td>
<td>6.86</td>
<td>12.68</td>
</tr>
<tr>
<td>TOGO</td>
<td>▼</td>
<td>5</td>
<td>6.27</td>
<td>17.34</td>
</tr>
<tr>
<td>CAMEROON</td>
<td>▼</td>
<td>5</td>
<td>6.93</td>
<td>15.47</td>
</tr>
<tr>
<td>NIGERIA</td>
<td>▼</td>
<td>7</td>
<td>6.88</td>
<td>17.46</td>
</tr>
<tr>
<td>GABON</td>
<td>▼</td>
<td>8</td>
<td>8.19</td>
<td>16.97</td>
</tr>
</tbody>
</table>

Note: Missing Data
<table>
<thead>
<tr>
<th>country</th>
<th>polityiv</th>
<th>f-house</th>
<th>log-gdppc</th>
<th>primary</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUKINA FASO</td>
<td></td>
<td>6</td>
<td>6.23</td>
<td>5.92</td>
</tr>
<tr>
<td>LIBERIA</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIERRA LEONE</td>
<td></td>
<td>3</td>
<td>6.60</td>
<td></td>
</tr>
<tr>
<td>GHANA</td>
<td></td>
<td>6</td>
<td>6.86</td>
<td>12.68</td>
</tr>
<tr>
<td>TOGO</td>
<td></td>
<td>5</td>
<td>6.27</td>
<td>17.34</td>
</tr>
<tr>
<td>CAMEROON</td>
<td></td>
<td>5</td>
<td>6.93</td>
<td>15.47</td>
</tr>
<tr>
<td>NIGERIA</td>
<td></td>
<td>7</td>
<td>6.88</td>
<td>17.46</td>
</tr>
<tr>
<td>GABON</td>
<td></td>
<td>8</td>
<td>8.19</td>
<td>16.97</td>
</tr>
</tbody>
</table>

Stewart (Princeton)
Run whatever analysis model you wanted to run.
Run whatever analysis model you wanted to run. 

\( \times 5 \)
But how does it work?
Let’s look at the extreme case first.
Missing Data

Stewart (Princeton)

Mar 27-Apr 5, 2017 150 / 200
Stewart (Princeton)
ARBITRARY PATTERNS OF MISMEASUREMENT & MISSINGNESS:

<table>
<thead>
<tr>
<th>Country</th>
<th>Polity</th>
<th>GDPpc</th>
<th>Pop</th>
<th>Expected GDPpc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burkina Faso</td>
<td>4</td>
<td>6.23</td>
<td>5.92</td>
<td>5.92</td>
</tr>
<tr>
<td>Liberia</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Sierra Leone</td>
<td>3</td>
<td>6.60</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Ghana</td>
<td>≈9</td>
<td>6</td>
<td>6.86</td>
<td>12.68</td>
</tr>
<tr>
<td>Togo</td>
<td>NA</td>
<td>6.27</td>
<td>17.34</td>
<td></td>
</tr>
<tr>
<td>Cameroon</td>
<td>≈6</td>
<td>5</td>
<td>6.93</td>
<td>15.47</td>
</tr>
<tr>
<td>Nigeria</td>
<td>≈7</td>
<td>6.88</td>
<td>17.46</td>
<td></td>
</tr>
<tr>
<td>Gabon</td>
<td>≈8</td>
<td>8.19</td>
<td>16.97</td>
<td></td>
</tr>
</tbody>
</table>
ARBITRARY PATTERNS OF MISMEASUREMENT & MISSINGNESS:

<table>
<thead>
<tr>
<th>country</th>
<th>polityiv</th>
<th>f-house</th>
<th>log-gdppc</th>
<th>primary</th>
</tr>
</thead>
<tbody>
<tr>
<td>BUKINA FASO</td>
<td>4</td>
<td>≈6</td>
<td>6.23</td>
<td>5.92</td>
</tr>
<tr>
<td>LIBERIA</td>
<td>NA</td>
<td>3</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>SIERRA LEONE</td>
<td>3</td>
<td>3</td>
<td>6.60</td>
<td>NA</td>
</tr>
<tr>
<td>GHANA</td>
<td>≈9</td>
<td>6</td>
<td>6.86</td>
<td>12.68</td>
</tr>
<tr>
<td>TOGO</td>
<td>NA</td>
<td>5</td>
<td>6.27</td>
<td>17.34</td>
</tr>
<tr>
<td>CAMEROON</td>
<td>≈6</td>
<td>5</td>
<td>6.93</td>
<td>15.47</td>
</tr>
<tr>
<td>NIGERIA</td>
<td>≈5</td>
<td>7</td>
<td>≈6.88</td>
<td>17.46</td>
</tr>
<tr>
<td>GABON</td>
<td>≈6</td>
<td>8</td>
<td>≈8.19</td>
<td>≈16.97</td>
</tr>
</tbody>
</table>
The Multiple Imputation Scheme
The Multiple Imputation Scheme

incomplete data
The Multiple Imputation Scheme

- incomplete data
- posterior predictive draws
- imputed datasets

Stewart (Princeton)
The Multiple Imputation Scheme

incomplete data

posterior predictive draws

imputed datasets

analysis

separate results
The Multiple Imputation Scheme

- **incomplete data**
- **posterior predictive draws**
- **imputed datasets**
- **analysis**
- **separate results**
- **combination**
- **final results**
Mismeasured at random (MMAR).
You have to know how things were mismeasured.
You have to know $f(x_i|x_i^*)$. 

(2)
(3*)

Measurement error and ideal data are statistically dual.
OUR SPECIFIC MODEL
OUR SPECIFIC MODEL

\[ (y_i, x_i^*) \sim \mathcal{MVN}(\mu, \Sigma) \]
OUR SPECIFIC MODEL

ideal data

\((y_i, x_i^*) \sim \mathcal{MN}(\mu, \Sigma)\)

\(x_i \sim \mathcal{N}(x_i^*, \sigma_u^2)\)

measurement error
EM
EM

(\hat{\mu}, \hat{\Sigma})
$EM$ \\
\[ (\hat{\mu}, \hat{\Sigma}) \] \\
\[ x^*_\text{imp} \]
EM

$(\hat{\mu}, \hat{\Sigma})$

$x_{imp}^*, y_{imp}^{mis}$
CHOOSE A VALUE OF $\sigma_u^2$
CHOOSE A VALUE OF $\sigma^2_u$

fully observed     my variable     fully missing
Some simulations.
Slope

0.0
0.2
0.4
0.6
0.8
1.0

0.0
0.5
1.0
1.5
2.0
2.5
3.0

fully observed

assumed amount of error

Stewart (Princeton)

Mar 27-Apr 5, 2017 172 / 200

Missing Data
true amount of ME

fully observed

assumed amount of error

fully missing

Stewart (Princeton)

Missing Data

Mar 27-Apr 5, 2017 174 / 200
true amount of ME

true slope

assumed amount of error

fully observed

fully missing

Stewart (Princeton)

Mar 27-Apr 5, 2017 175 / 200
The graph illustrates the relationship between the assumed amount of error and the true amount of ME. The x-axis represents the assumed amount of error, ranging from 0.0 to 1.0. The y-axis represents the slope, ranging from 0.0 to 3.0. The graph shows that as the assumed amount of error increases, the true amount of ME remains constant. The dashed line indicates the infeasible estimator, and the solid line shows the fully observed and fully missing conditions.
true amount of ME

infeasible estimator

denial estimator

fully observed

assumed amount of error

fully missing
Missing Data

- Infeasible estimator
- Overimputation
- Denial estimator

True amount of ME

Assumed amount of error
Slope

0.0  0.2  0.4  0.6  0.8  1.0

0.0  0.5  1.0  1.5  2.0  2.5  3.0

true amount of ME

infeasible estimator

MO at the true amount

overimputation

denial estimator

fully observed  assumed amount of error  fully missing

Stewart (Princeton)
CHOOSE A RANGE OF $\sigma^2_u$

Fully observed

my variable

Fully missing
true amount of ME

overimputation bounds

denial estimator

true slope

MO at the true amount

fully observed

assumed amount of error

fully missing
Missing Data  | Measurement Error

| 20 Years Ago |  
|--------------|------------------|

<table>
<thead>
<tr>
<th>Today</th>
<th></th>
</tr>
</thead>
</table>

TAILORED METHODS:
- Model dependent
- Difficult to implement
- Dubious assumptions

MULTIPLE IMPUTATION:
- Broadly applicable
- Easy to implement
- Widely used.

Stewart (Princeton)
<table>
<thead>
<tr>
<th>20 Years Ago</th>
<th>Today</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Missing Data</strong></td>
<td><strong>Measurement Error</strong></td>
</tr>
<tr>
<td>TAILORED METHODS:</td>
<td>TAILORED METHODS:</td>
</tr>
</tbody>
</table>

**Model dependent**
- Difficult to implement
- Dubious assumptions

**MULTIPLE IMPUTATION**
- Broadly applicable
- Easy to implement
- Widely used.

Stewart (Princeton)
<table>
<thead>
<tr>
<th></th>
<th>Missing Data</th>
<th>Measurement Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>20 Years Ago</strong></td>
<td><strong>TAILLORED METHODS:</strong> Model dependent</td>
<td><strong>TAILLORED METHODS:</strong> Model dependent</td>
</tr>
<tr>
<td><strong>Today</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Missing Data</td>
<td>Measurement Error</td>
</tr>
<tr>
<td>----------------</td>
<td>--------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>20 Years Ago</td>
<td>TAILORED METHODS: Model dependent Difficult to implement</td>
<td>TAILORED METHODS: Model dependent Difficult to implement</td>
</tr>
<tr>
<td>Today</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 Years Ago</td>
<td>Today</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-------</td>
<td></td>
</tr>
<tr>
<td><strong>Missing Data</strong></td>
<td><strong>Measurement Error</strong></td>
<td></td>
</tr>
<tr>
<td>TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions</td>
<td>TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions</td>
<td></td>
</tr>
</tbody>
</table>
### Missing Data

<table>
<thead>
<tr>
<th>20 Years Ago</th>
<th>Today</th>
</tr>
</thead>
</table>
| **TAILORED METHODS:**
  Model dependent
  Difficult to implement
  Dubious assumptions |
| **MULTIPLE IMPUTATION:** |

### Measurement Error

<table>
<thead>
<tr>
<th>20 Years Ago</th>
<th>Today</th>
</tr>
</thead>
</table>
| **TAILORED METHODS:**
  Model dependent
  Difficult to implement
  Dubious assumptions |
|              |       |
### Missing Data

<table>
<thead>
<tr>
<th>20 Years Ago</th>
<th>Today</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions</td>
<td>MULTIPLE IMPUTATION: Broadly applicable</td>
</tr>
</tbody>
</table>

### Measurement Error

<table>
<thead>
<tr>
<th>20 Years Ago</th>
<th>Today</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Missing Data</td>
</tr>
<tr>
<td>----------------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td><strong>20 Years Ago</strong></td>
<td><strong>TAILORED METHODS:</strong></td>
</tr>
<tr>
<td></td>
<td>Model dependent</td>
</tr>
<tr>
<td></td>
<td>Difficult to implement</td>
</tr>
<tr>
<td></td>
<td>Dubious assumptions</td>
</tr>
<tr>
<td><strong>Today</strong></td>
<td><strong>MULTIPLE IMPUTATION:</strong></td>
</tr>
<tr>
<td></td>
<td>Broadly applicable</td>
</tr>
<tr>
<td></td>
<td>Easy to implement</td>
</tr>
<tr>
<td>Missing Data</td>
<td>Measurement Error</td>
</tr>
<tr>
<td>-------------------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td><strong>20 Years Ago</strong></td>
<td><strong>TAILORED METHODS:</strong></td>
</tr>
<tr>
<td></td>
<td>Model dependent</td>
</tr>
<tr>
<td></td>
<td>Difficult to implement</td>
</tr>
<tr>
<td></td>
<td>Dubious assumptions</td>
</tr>
<tr>
<td><strong>Today</strong></td>
<td><strong>TAILORED METHODS:</strong></td>
</tr>
<tr>
<td></td>
<td>Model dependent</td>
</tr>
<tr>
<td></td>
<td>Difficult to implement</td>
</tr>
<tr>
<td></td>
<td>Dubious assumptions</td>
</tr>
<tr>
<td></td>
<td><strong>MULTIPLE IMPUTATION:</strong></td>
</tr>
<tr>
<td></td>
<td>Broadly applicable</td>
</tr>
<tr>
<td></td>
<td>Easy to implement</td>
</tr>
<tr>
<td></td>
<td>Widely used.</td>
</tr>
<tr>
<td>Missing Data</td>
<td>Measurement Error</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------------</td>
</tr>
<tr>
<td><strong>20 Years Ago</strong></td>
<td><strong>Today</strong></td>
</tr>
<tr>
<td>TAILORED METHODS: Model dependent</td>
<td>TAILORED METHODS: Model dependent</td>
</tr>
<tr>
<td>Difficult to implement</td>
<td>Difficult to implement</td>
</tr>
<tr>
<td>Dubious assumptions</td>
<td>Dubious assumptions</td>
</tr>
<tr>
<td>MULTIPLE IMPUTATION:</td>
<td>MULTIPLE OVERIMPUTATION</td>
</tr>
<tr>
<td>Broadly applicable</td>
<td></td>
</tr>
<tr>
<td>Easy to implement</td>
<td></td>
</tr>
<tr>
<td>Widely used.</td>
<td></td>
</tr>
</tbody>
</table>
1. Mixture Models
   - Basic Mixtures
   - Application: Mixtures as Preprocessing
   - Application: Mixture of Regressions

2. Expectation Maximization
   - EM for Probit Regression
   - EM for Gaussian Mixtures
   - EM in General

3. Missing Data
   - Motivating Example
   - Overview and Assumptions
   - Existing Heuristics
   - Application Specific Approaches
   - Multiple Imputation
   - The Full Amelia Scheme

4. Measurement Error

5. Appendix: Additional Details and Examples
Mixture Models
- Basic Mixtures
- Application: Mixtures as Preprocessing
- Application: Mixture of Regressions

Expectation Maximization
- EM for Probit Regression
- EM for Gaussian Mixtures
- EM in General

Missing Data
- Motivating Example
- Overview and Assumptions
- Existing Heuristics
- Application Specific Approaches
- Multiple Imputation
- The Full Amelia Scheme

Measurement Error

Appendix: Additional Details and Examples
Goal: estimate $\beta_1$, where $X_2$ has $\lambda$ missing values ($y, X_1$ are fully observed).

$E(y) = X_1 \beta_1 + X_2 \beta_2$

The choice in real research:

- Infeasible Estimator: Regress $y$ on $X_1$ and a fully observed $X_2$, and use $b_{I1}$, the coefficient on $X_1$.

- Omitted Variable Estimator: Omit $X_2$ and estimate $\beta_1$ by $b_{O1}$, the slope from regressing $y$ on $X_1$.

- Listwise Deletion Estimator: Perform listwise deletion on \{y, X_1, X_2\}, and then estimate $\beta_1$ as $b_{L1}$, the coefficient on $X_1$ when regressing $y$ on $X_1$ and $X_2$. 
**Goal:** estimate $\beta_1$, where $X_2$ has $\lambda$ missing values ($y, X_1$ are fully observed).
How Bad Is Listwise Deletion?

**Goal:** estimate $\beta_1$, where $X_2$ has $\lambda$ missing values ($y$, $X_1$ are fully observed).

$$E(y) = X_1\beta_1 + X_2\beta_2$$
How Bad Is Listwise Deletion?

**Goal:** estimate $\beta_1$, where $X_2$ has $\lambda$ missing values ($y$, $X_1$ are fully observed).

$$E(y) = X_1\beta_1 + X_2\beta_2$$

The choice in real research:
How Bad Is Listwise Deletion?

**Goal:** estimate $\beta_1$, where $X_2$ has $\lambda$ missing values ($y$, $X_1$ are fully observed).

$$E(y) = X_1\beta_1 + X_2\beta_2$$

**The choice in real research:**

**Infeasible Estimator** Regress $y$ on $X_1$ and a fully observed $X_2$, and use $b_1^l$, the coefficient on $X_1$. 
How Bad Is Listwise Deletion?

**Goal:** estimate $\beta_1$, where $X_2$ has $\lambda$ missing values ($y$, $X_1$ are fully observed).

$$E(y) = X_1\beta_1 + X_2\beta_2$$

**The choice in real research:**

**Infeasible Estimator** Regress $y$ on $X_1$ and a fully observed $X_2$, and use $b_1^I$, the coefficient on $X_1$.

**Omitted Variable Estimator** Omit $X_2$ and estimate $\beta_1$ by $b_1^O$, the slope from regressing $y$ on $X_1$. 

Stewart (Princeton)
How Bad Is Listwise Deletion?

**Goal:** estimate $\beta_1$, where $X_2$ has $\lambda$ missing values ($y$, $X_1$ are fully observed).

$$E(y) = X_1\beta_1 + X_2\beta_2$$

**The choice in real research:**

**Infeasible Estimator**
Regress $y$ on $X_1$ and a fully observed $X_2$, and use $b_1^I$, the coefficient on $X_1$.

**Omitted Variable Estimator**
Omit $X_2$ and estimate $\beta_1$ by $b_1^O$, the slope from regressing $y$ on $X_1$.

**Listwise Deletion Estimator**
Perform listwise deletion on \{y, X_1, X_2\}, and then estimate $\beta_1$ as $b_1^L$, the coefficient on $X_1$ when regressing $y$ on $X_1$ and $X_2$. 
In the best case scenario for listwise deletion (MCAR), should we delete listwise or omit the variable?

Mean Square Error as a measure of the badness of an estimator \( \hat{a} \) of \( a \).
In the **best** case scenerio for listwise deletion (MCAR), should we delete listwise or omit the variable?

Mean Square Error as a measure of the badness of an estimator \( \hat{a} \) of \( a \).

\[
\text{MSE}(\hat{a}) = E[(\hat{a} - a)^2]
\]
In the best case scenario for listwise deletion (MCAR), should we delete listwise or omit the variable?

Mean Square Error as a measure of the badness of an estimator \( \hat{a} \) of \( a \).

\[
\text{MSE}(\hat{a}) = E[(\hat{a} - a)^2] \\
= V(\hat{a}) + [E(\hat{a} - a)]^2
\]
In the best case scenario for listwise deletion (MCAR), should we delete listwise or omit the variable?

Mean Square Error as a measure of the badness of an estimator $\hat{a}$ of $a$.

$$\text{MSE}(\hat{a}) = E[(\hat{a} - a)^2]$$
$$= V(\hat{a}) + [E(\hat{a} - a)]^2$$
$$= \text{Variance}(\hat{a}) + \text{bias}(\hat{a})^2$$
In the **best** case scenario for listwise deletion (MCAR), should we delete listwise or omit the variable?

Mean Square Error as a measure of the badness of an estimator $\hat{a}$ of $a$.

$$\text{MSE}(\hat{a}) = E[(\hat{a} - a)^2]$$

$$= V(\hat{a}) + [E(\hat{a} - a)]^2$$

$$= \text{Variance}(\hat{a}) + \text{bias}(\hat{a})^2$$

To compare, compute
In the best case scenario for listwise deletion (MCAR), should we delete listwise or omit the variable?

Mean Square Error as a measure of the badness of an estimator $\hat{a}$ of $a$.

$$\text{MSE}(\hat{a}) = E[(\hat{a} - a)^2]$$

$$= V(\hat{a}) + [E(\hat{a} - a)]^2$$

$$= \text{Variance}(\hat{a}) + \text{bias}(\hat{a})^2$$

To compare, compute

$$\text{MSE}(b_1^L) - \text{MSE}(b_1^O) =$$
In the best case scenario for listwise deletion (MCAR), should we delete listwise or omit the variable?

Mean Square Error as a measure of the badness of an estimator $\hat{a}$ of $a$.

$$\text{MSE}(\hat{a}) = E[(\hat{a} - a)^2]$$
$$= V(\hat{a}) + [E(\hat{a} - a)]^2$$
$$= \text{Variance}(\hat{a}) + \text{bias}(\hat{a})^2$$

To compare, compute

$$\text{MSE}(b^{L}_1) - \text{MSE}(b^{O}_1) = \begin{cases} > 0 & \text{when omitting the variable is better} \\ < 0 & \text{when listwise deletion is better} \end{cases}$$
Derivation and Implications

\[
\text{MSE}(b_{L1}) - \text{MSE}(b_{O1}) = (\lambda_1 - \lambda V(b_{I1})) + F[V(b_{I2}) - \beta_2 \beta'_{2}]
\]

1. Missingness part (>0) is an extra tilt away from listwise deletion
2. Observed part is the standard bias-efficiency tradeoff of omitting variables, even without missingness
3. How big is $\lambda$ usually? (from literature review in King et al 2001)
   ▶ $\lambda \approx \frac{1}{3}$ on average in real political science articles
   ▶ $> \frac{1}{2}$ at the Polmeth Conference
   ▶ Larger for authors who work harder to avoid omitted variable bias

Stewart (Princeton)
Derivation and Implications

$$\text{MSE}(b_1^L) - \text{MSE}(b_1^O)$$
Derivation and Implications

\[
\text{MSE}(b_1^L) - \text{MSE}(b_1^O) = \left( \frac{\lambda}{1 - \lambda} V(b_1^I) \right) + F[V(b_2^I) - \beta_2'\beta_2] F'
\]

1. Missingness part \( > 0 \) is an extra tilt away from listwise deletion
2. Observed part is the standard bias-efficiency tradeoff of omitting variables, even without missingness
3. How big is \( \lambda \) usually? (from literature review in King et al 2001)
   - \( \lambda \approx 1/3 \) on average in real political science articles
   - \( > 1/2 \) at the Polmeth Conference
   - Larger for authors who work harder to avoid omitted variable bias
Derivation and Implications

\[
\text{MSE}(b^L_1) - \text{MSE}(b^O_1) = \left( \frac{\lambda}{1 - \lambda} V(b^l_1) \right) + F[V(b^l_2) - \beta_2 \beta'_2] F' \\
= (\text{Missingness part}) + (\text{Observed part})
\]
Derivation and Implications

\[ \text{MSE}(b_1^L) - \text{MSE}(b_1^O) = \left( \frac{\lambda}{1 - \lambda} \text{V}(b_1^I) \right) + F[\text{V}(b_2^I) - \beta_2\beta'_2]F' \]

\[ = \text{(Missingness part)} + \text{(Observed part)} \]

1. Missingness part \((>0)\) is an extra tilt away from listwise deletion
Derivation and Implications

\[
\text{MSE}(b^L_1) - \text{MSE}(b^O_1) = \left( \frac{\lambda}{1 - \lambda} V(b^I_1) \right) + F[V(b^I_2) - \beta_2 \beta'_2] F' \\
= (\text{Missingness part}) + (\text{Observed part})
\]

1. Missingness part (> 0) is an extra tilt away from listwise deletion
2. Observed part is the standard bias-efficiency tradeoff of omitting variables, even without missingness
Derivation and Implications

\[
\text{MSE}(b_1^L) - \text{MSE}(b_1^O) = \left( \frac{\lambda}{1 - \lambda} \text{V}(b_1^l) \right) + F[\text{V}(b_2^l) - \beta_2\beta'_2]F'
\]

= (Missingness part) + (Observed part)

1. Missingness part \((> 0)\) is an extra tilt away from listwise deletion
2. Observed part is the standard bias-efficiency tradeoff of omitting variables, even without missingness
3. How big is \(\lambda\) usually? (from literature review in King et al 2001)
Derivation and Implications

\[
\text{MSE}(b^L_1) - \text{MSE}(b^O_1) = \left( \frac{\lambda}{1 - \lambda} V(b^l_1) \right) + F[V(b^l_2) - \beta_2 \beta'_2]F' 
\]

\[= (\text{Missingness part}) + (\text{Observed part})\]

1. Missingness part \((> 0)\) is an extra tilt away from listwise deletion
2. Observed part is the standard bias-efficiency tradeoff of omitting variables, even without missingness
3. How big is \(\lambda\) usually? \((\text{from literature review in King et al 2001})\)
   - \(\lambda \approx 1/3\) on average in real political science articles
Derivation and Implications

\[
\text{MSE}(b_1^L) - \text{MSE}(b_1^O) = \left( \frac{\lambda}{1 - \lambda} \text{V}(b_1') \right) + F[V(b_2') - \beta_2\beta_2']F' \\
= (\text{Missingness part}) + (\text{Observed part})
\]

1. Missingness part \((> 0)\) is an extra tilt away from listwise deletion
2. Observed part is the standard bias-efficiency tradeoff of omitting variables, even without missingness
3. How big is \(\lambda\) usually? (from literature review in King et al 2001)
   - \(\lambda \approx 1/3\) on average in real political science articles
   - \(> 1/2\) at the Polmeth Conference
Derivation and Implications

\[
\text{MSE}(b_1^L) - \text{MSE}(b_1^O) = \left(\frac{\lambda}{1 - \lambda} \text{V}(b_1^I)\right) + F[V(b_2^I) - \beta_2\beta_2']F' \\
= \text{(Missingness part)} + \text{(Observed part)}
\]

1. Missingness part \((> 0)\) is an extra tilt away from listwise deletion
2. Observed part is the standard bias-efficiency tradeoff of omitting variables, even without missingness
3. How big is \(\lambda\) usually? (from literature review in King et al 2001)
   - \(\lambda \approx 1/3\) on average in real political science articles
   - \(> 1/2\) at the Polmeth Conference
   - Larger for authors who work harder to avoid omitted variable bias
4. If $\lambda \approx 0.5$, the contribution of the missingness (tilting away from choosing listwise deletion over omitting variables) is

$\text{RMSE difference} = \sqrt{\lambda V(b_I)} = \sqrt{0.5 - 0.5SE(b_I)} = SE(b_I)$

(The sqrt of only one piece, for simplicity, not the difference.)

5. Result: The point estimate in the average political science article is about an additional standard error farther away from the truth because of listwise deletion (as compared to omitting $X_2$ entirely).

6. Conclusion: Listwise deletion is often as bad a problem as the much better known omitted variable bias — in the best case scenario (MCAR).
4. If $\lambda \approx 0.5$, the contribution of the missingness (tilting away from choosing listwise deletion over omitting variables) is

\[ \text{RMSE difference} = \sqrt{\frac{\lambda}{1 - \lambda} V(b_1')} = \sqrt{\frac{0.5}{1 - 0.5}} \text{SE}(b_1') = \text{SE}(b_1') \]
Derivation and Implications

4. If $\lambda \approx 0.5$, the contribution of the missingness (tilting away from choosing listwise deletion over omitting variables) is

$$\text{RMSE difference} = \sqrt{\frac{\lambda}{1-\lambda} V(b_1')} = \sqrt{\frac{0.5}{1-0.5}} \text{SE}(b_1') = \text{SE}(b_1')$$

(The sqrt of only one piece, for simplicity, not the difference.)
4. If \( \lambda \approx 0.5 \), the contribution of the missingness (tilting away from choosing listwise deletion over omitting variables) is

\[
\text{RMSE difference} = \sqrt{\frac{\lambda}{1 - \lambda}} V(b_1^l) = \sqrt{\frac{0.5}{1 - 0.5}} \text{SE}(b_1^l) = \text{SE}(b_1^l)
\]

(The sqrt of only one piece, for simplicity, not the difference.)

5. **Result:** The point estimate in the average political science article is about an additional standard error farther away from the truth because of listwise deletion (as compared to omitting \( X_2 \) entirely).
Derivation and Implications

4. If $\lambda \approx 0.5$, the contribution of the missingness (tilting away from choosing listwise deletion over omitting variables) is

$$\text{RMSE difference} = \sqrt{\frac{\lambda}{1-\lambda} V(b_1)} = \sqrt{\frac{0.5}{1-0.5}} \text{SE}(b_1) = \text{SE}(b_1)$$

(The sqrt of only one piece, for simplicity, not the difference.)

5. Result: The point estimate in the average political science article is about an additional standard error farther away from the truth because of listwise deletion (as compared to omitting $X_2$ entirely).

6. Conclusion: Listwise deletion is often as bad a problem as the much better known omitted variable bias — in the best case scenario (MCAR)
The Best Case for Listwise Deletion

Listwise deletion is better than MI when all 4 hold:

1. The analysis model is conditional on \( X \) (like regression) and functional form is correct (so listwise deletion is consistent and the characteristic robustness of regression is not lost when applied to data with slight measurement error, endogeneity, nonlinearity, etc.).

2. NI missingness in \( X \) and no external variables are available that could be used in an imputation stage to fix the problem.

3. Missingness in \( X \) is not a function of \( Y \)

4. The \( n \) left after listwise deletion is so large that the efficiency loss does not counter balance the biases induced by the other conditions. I.e., you don’t trust data to impute \( D_{mis} \) but still trust it to analyze \( D_{obs} \)
The Best Case for Listwise Deletion

Listwise deletion is better than MI when all 4 hold:
The Best Case for Listwise Deletion

Listwise deletion is better than MI when all 4 hold:

1. The analysis model is conditional on $X$ (like regression) and functional form is correct (so listwise deletion is consistent and the characteristic robustness of regression is not lost when applied to data with slight measurement error, endogeneity, nonlinearity, etc.).
The Best Case for Listwise Deletion

Listwise deletion is better than MI when all 4 hold:

1. The analysis model is conditional on $X$ (like regression) and functional form is correct (so listwise deletion is consistent and the characteristic robustness of regression is not lost when applied to data with slight measurement error, endogeneity, nonlinearity, etc.).

2. NI missingness in $X$ and no external variables are available that could be used in an imputation stage to fix the problem.
The Best Case for Listwise Deletion

Listwise deletion is better than MI when all 4 hold:

1. The analysis model is conditional on $X$ (like regression) and functional form is correct (so listwise deletion is consistent and the characteristic robustness of regression is not lost when applied to data with slight measurement error, endogeneity, nonlinearity, etc.).

2. NI missingness in $X$ and no external variables are available that could be used in an imputation stage to fix the problem.

3. Missingness in $X$ is not a function of $Y$
The Best Case for Listwise Deletion

Listwise deletion is better than MI when all 4 hold:

1. The analysis model is conditional on $X$ (like regression) and functional form is correct (so listwise deletion is consistent and the characteristic robustness of regression is not lost when applied to data with slight measurement error, endogeneity, nonlinearity, etc.).

2. NI missingness in $X$ and no external variables are available that could be used in an imputation stage to fix the problem.

3. Missingness in $X$ is not a function of $Y$

4. The $n$ left after listwise deletion is so large that the efficiency loss does not counterbalance the biases induced by the other conditions.
The Best Case for Listwise Deletion

Listwise deletion is better than MI when all 4 hold:

1. The analysis model is conditional on $X$ (like regression) and functional form is correct (so listwise deletion is consistent and the characteristic robustness of regression is not lost when applied to data with slight measurement error, endogeneity, nonlinearity, etc.).

2. NI missingness in $X$ and no external variables are available that could be used in an imputation stage to fix the problem.

3. Missingness in $X$ is not a function of $Y$

4. The $n$ left after listwise deletion is so large that the efficiency loss does not counter balance the biases induced by the other conditions.

I.e., you don’t trust data to impute $D_{mis}$ but still trust it to analyze $D_{obs}$
Root Mean Square Error Comparisons

Each point is RMSE averaged over two regression coefficients in each of 100 simulated data sets. (IP and EMis have the same RMSE, which is lower than listwise deletion and higher than the complete data; its the same for EMB.)
Detailed Example: Support for Perot

1. Research question: were voters who did not share in the economic recovery more likely to support Perot in the 1996 presidential election?

2. Analysis model: linear regression

3. Data: 1996 National Election Survey (n=1714)

4. Dependent variable: Perot Feeling Thermometer

5. Key explanatory variables: retrospective and prospective evaluations of national economic performance and personal financial circumstances

6. Control variables: age, education, family income, race, gender, union membership, ideology

7. Extra variables included in the imputation model to help prediction: attention to the campaign; feeling thermometers for Clinton, Dole, Democrats, Republicans; PID; Partisan moderation; vote intention; marital status; Hispanic; party contact, number of organizations R is a paying member of, and active member of.

8. Include nonlinear terms: age$^2$
Detailed Example: Support for Perot

1. Research question: were voters who did not share in the economic recovery more likely to support Perot in the 1996 presidential election?
Detailed Example: Support for Perot

1. Research question: were voters who did not share in the economic recovery more likely to support Perot in the 1996 presidential election?

2. Analysis model: linear regression

3. Data: 1996 National Election Survey (n=1714)

4. Dependent variable: Perot Feeling Thermometer

5. Key explanatory variables: retrospective and prospective evaluations of national economic performance and personal financial circumstances

6. Control variables: age, education, family income, race, gender, union membership, ideology

7. Extra variables included in the imputation model to help prediction: attention to the campaign; feeling thermometers for Clinton, Dole, Democrats, Republicans; PID; Partisan moderation; vote intention; martial status; Hispanic; party contact, number of organizations R is a paying member of, and active member of.

8. Include nonlinear terms: age $^2$
Detailed Example: Support for Perot

1. Research question: were voters who did not share in the economic recovery more likely to support Perot in the 1996 presidential election?

2. Analysis model: linear regression

3. Data: 1996 National Election Survey (n=1714)
Detailed Example: Support for Perot

1. Research question: were voters who did not share in the economic recovery more likely to support Perot in the 1996 presidential election?
2. Analysis model: linear regression
3. Data: 1996 National Election Survey (n=1714)
4. Dependent variable: Perot Feeling Thermometer
Detailed Example: Support for Perot

1. Research question: were voters who did not share in the economic recovery more likely to support Perot in the 1996 presidential election?
2. Analysis model: linear regression
3. Data: 1996 National Election Survey (n=1714)
4. Dependent variable: Perot Feeling Thermometer
5. Key explanatory variables: retrospective and prospective evaluations of national economic performance and personal financial circumstances
Detailed Example: Support for Perot

1. Research question: were voters who did not share in the economic recovery more likely to support Perot in the 1996 presidential election?
2. Analysis model: linear regression
3. Data: 1996 National Election Survey (n=1714)
4. Dependent variable: Perot Feeling Thermometer
5. Key explanatory variables: retrospective and prospective evaluations of national economic performance and personal financial circumstances
6. Control variables: age, education, family income, race, gender, union membership, ideology

Extra variables included in the imputation model to help prediction:
- attention to the campaign
- feeling thermometers for Clinton, Dole, Democrats, Republicans
- PID
- Partisan moderation
- vote intention
- martial status
- Hispanic
- party contact, number of organizations R is a paying member of, and active member of.

Include nonlinear terms: age $^2$
Detailed Example: Support for Perot

1. Research question: were voters who did not share in the economic recovery more likely to support Perot in the 1996 presidential election?
2. Analysis model: linear regression
3. Data: 1996 National Election Survey (n=1714)
4. Dependent variable: Perot Feeling Thermometer
5. Key explanatory variables: retrospective and prospective evaluations of national economic performance and personal financial circumstances
6. Control variables: age, education, family income, race, gender, union membership, ideology
7. Extra variables included in the imputation model to help prediction: attention to the campaign; feeling thermometers for Clinton, Dole, Democrats, Republicans; PID; Partisan moderation; vote intention; marital status; Hispanic; party contact, number of organizations R is a paying member of, and active member of.
Detailed Example: Support for Perot

1. Research question: were voters who did not share in the economic recovery more likely to support Perot in the 1996 presidential election?

2. Analysis model: linear regression

3. Data: 1996 National Election Survey (n=1714)

4. Dependent variable: Perot Feeling Thermometer

5. Key explanatory variables: retrospective and prospective evaluations of national economic performance and personal financial circumstances

6. Control variables: age, education, family income, race, gender, union membership, ideology

7. Extra variables included in the imputation model to help prediction: attention to the campaign; feeling thermometers for Clinton, Dole, Democrats, Republicans; PID; Partisan moderation; vote intention; martial status; Hispanic; party contact, number of organizations R is a paying member of, and active member of.

8. Include nonlinear terms: age^2
9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.

10. Run Amelia to generate 5 imputed data sets.

11. Key substantive result is the coefficient on retrospective economic evaluations (ranges from 1 to 5):

- Listwise deletion: 43.90
- Multiple imputation: 43.65

so \((5 - 1) \times 1.65 = 6.6\), which is also a percentage of the range of \(Y\).

(a) MI estimator is more efficient, with a smaller SE
(b) The MI estimator is 4 times larger
(c) Based on listwise deletion, there is no evidence that perception of poor economic performance is related to support for Perot
(d) Based on the MI estimator, R's with negative retrospective economic evaluations are more likely to have favorable views of Perot.

Stewart (Princeton)
9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.

10. Run Amelia to generate 5 imputed data sets.
9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.

10. Run Amelia to generate 5 imputed data sets.

11. Key substantive result is the coefficient on retrospective economic evaluations (ranges from 1 to 5):
9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.
10. Run Amelia to generate 5 imputed data sets.
11. Key substantive result is the coefficient on retrospective economic evaluations (ranges from 1 to 5):

<table>
<thead>
<tr>
<th>Method</th>
<th>Coefficient</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listwise deletion</td>
<td>.43</td>
<td>(.90)</td>
</tr>
<tr>
<td>Multiple imputation</td>
<td>1.65</td>
<td>(.72)</td>
</tr>
</tbody>
</table>
9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.

10. Run Amelia to generate 5 imputed data sets.

11. Key substantive result is the coefficient on retrospective economic evaluations (ranges from 1 to 5):

<table>
<thead>
<tr>
<th>Method</th>
<th>Coefficient</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listwise deletion</td>
<td>.43</td>
<td>(.90)</td>
</tr>
<tr>
<td>Multiple imputation</td>
<td>1.65</td>
<td>(.72)</td>
</tr>
</tbody>
</table>

so \((5 - 1) \times 1.65 = 6.6\), which is also a percentage of the range of \(Y\).
9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.

10. Run Amelia to generate 5 imputed data sets.

11. Key substantive result is the coefficient on retrospective economic evaluations (ranges from 1 to 5):

- Listwise deletion: 0.43 (.90)
- Multiple imputation: 1.65 (.72)

so \((5 - 1) \times 1.65 = 6.6\), which is also a percentage of the range of \(Y\).

(a) MI estimator is more efficient, with a smaller SE
9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.

10. Run Amelia to generate 5 imputed data sets.

11. Key substantive result is the coefficient on retrospective economic evaluations (ranges from 1 to 5):

   Listwise deletion: 0.43 (.90)
   Multiple imputation: 1.65 (.72)

   So $(5 - 1) \times 1.65 = 6.6$, which is also a percentage of the range of $Y$.

   (a) MI estimator is more efficient, with a smaller SE
   (b) The MI estimator is 4 times larger
9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.

10. Run Amelia to generate 5 imputed data sets.

11. Key substantive result is the coefficient on retrospective economic evaluations (ranges from 1 to 5):

   Listwise deletion \( \hat{\beta} = 0.43 \) 
   \( (0.90) \)

   Multiple imputation \( \hat{\beta} = 1.65 \) 
   \( (0.72) \)

so \((5 - 1) \times 1.65 = 6.6\), which is also a percentage of the range of \( Y \).

(a) MI estimator is more efficient, with a smaller SE

(b) The MI estimator is 4 times larger

(c) Based on listwise deletion, there is no evidence that perception of poor economic performance is related to support for Perot.
9. Transform variables to more closely approximate distributional assumptions: logged number of organizations participating in.

10. Run Amelia to generate 5 imputed data sets.

11. Key substantive result is the coefficient on retrospective economic evaluations (ranges from 1 to 5):

   Listwise deletion       0.43
                          (.90)

   Multiple imputation     1.65
                          (.72)

so \((5 - 1) \times 1.65 = 6.6\), which is also a percentage of the range of \(Y\).

(a) MI estimator is more efficient, with a smaller SE
(b) The MI estimator is 4 times larger
(c) Based on listwise deletion, there is no evidence that perception of poor economic performance is related to support for Perot
(d) Based on the MI estimator, R’s with negative retrospective economic evaluations are more likely to have favorable views of Perot.
Include: (1) fixed effects, (2) time trends, and (3) priors for cells

MI in Time Series Cross-Section Data

Include: (1) fixed effects, (2) time trends, and (3) priors for cells

MI in Time Series Cross-Section Data

Include: (1) fixed effects, (2) time trends, and (3) priors for cells
MI in Time Series Cross-Section Data

Include: (1) fixed effects, (2) time trends, and (3) priors for cells.
Read: James Honaker and Gary King, "What to do About Missing Values in Time Series Cross-Section Data,”
http://gking.harvard.edu/files/abs/pr-abs.shtml
Imputation one Observation at a time

Circles = true GDP; green = no time trends; blue = polynomials; red = LOESS
Priors on Cell Values

Recall:
\[ p(\theta | y) = p(\theta) \prod_{i=1}^{n} L_i(\theta | y) \]

Take logs:
\[ \ln p(\theta | y) = \ln p(\theta) + \sum_{i=1}^{n} \ln L_i(\theta | y) \]

Suppose prior is of the same form, \( p(\theta | y) = L_i(\theta | y) \); then its just another observation:
\[ \ln p(\theta | y) = \sum_{i=1}^{n+1} \ln L_i(\theta | y) \]

Honaker and King show how to modify these “data augmentation priors” to put priors on \( \mu \) and \( \sigma \) (or \( \beta \)).

Stewart (Princeton)
Priors on Cell Values

- Recall: \( p(\theta|y) = p(\theta) \prod_{i=1}^{n} L_i(\theta|y) \)
Priors on Cell Values

- Recall: \( p(\theta|y) = p(\theta) \prod_{i=1}^{n} L_i(\theta|y) \)
- Take logs: \( \ln p(\theta|y) = \ln[p(\theta)] + \sum_{i=1}^{n} \ln L_i(\theta|y) \)
Priors on Cell Values

- Recall: \( p(\theta|y) = p(\theta) \prod_{i=1}^{n} L_i(\theta|y) \)
- take logs: \( \ln p(\theta|y) = \ln[p(\theta)] + \sum_{i=1}^{n} \ln L_i(\theta|y) \)
- \( \implies \) Suppose prior is of the same form, \( p(\theta|y) = L_i(\theta|y) \); then its just another observation: \( \ln p(\theta|y) = \sum_{i=1}^{n+1} \ln L_i(\theta|y) \)
Priors on Cell Values

- Recall: \( p(\theta|y) = p(\theta) \prod_{i=1}^{n} L_i(\theta|y) \)
- take logs: \( \ln p(\theta|y) = \ln[p(\theta)] + \sum_{i=1}^{n} \ln L_i(\theta|y) \)
- Suppose prior is of the same form, \( p(\theta|y) = L_i(\theta|y) \); then its just another observation: \( \ln p(\theta|y) = \sum_{i=1}^{n+1} \ln L_i(\theta|y) \)
- Honaker and King show how to modify these “data augmentation priors” to put priors on missing values rather than on \( \mu \) and \( \sigma \) (or \( \beta \)).
Posterior imputation: mean=0, prior mean=5

Left column: holds prior $N(5, \lambda)$ constant ($\lambda = 1$) and changes predictive strength (the covariance, $\sigma_{12}$).

Right column: holds predictive strength of data constant (at $\sigma_{12} = 0.5$) and changes the strength of the prior ($\lambda$).
Posterior imputation: mean=0, prior mean=5

**Left column:** holds prior $N(5, \lambda)$ constant ($\lambda = 1$) and changes predictive strength (the covariance, $\sigma_{12}$).

**Right column:** holds predictive strength of data constant (at $\sigma_{12} = 0.5$) and changes the strength of the prior ($\lambda$).
Prior: $p(x_{12}) = N(5, \lambda)$. The parameter approaches the theoretical limits (dashed lines), upper bound is what is generated when the missing value is filled in with the expectation; lower bound is the parameter when the model is estimated without priors. The overall movement is small.
Replication of Baum and Lake; Imputation Model Fit

Barbados

Chile

Ghana

Iceland

Ecuador

Greece

Mongolia

Nepal

Black = observed. Blue circles = five imputations; Bars = 95% CIs
<table>
<thead>
<tr>
<th></th>
<th>Listwise Deletion</th>
<th>Multiple Imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Life Expectancy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rich Democracies</td>
<td>-.072</td>
<td>.233</td>
</tr>
<tr>
<td></td>
<td>(.179)</td>
<td>(.037)</td>
</tr>
<tr>
<td>Poor Democracies</td>
<td>-.082</td>
<td>.120</td>
</tr>
<tr>
<td></td>
<td>(.040)</td>
<td>(.099)</td>
</tr>
<tr>
<td>N</td>
<td>1789</td>
<td>5627</td>
</tr>
<tr>
<td><strong>Secondary Education</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rich Democracies</td>
<td>.948</td>
<td>.948</td>
</tr>
<tr>
<td></td>
<td>(.002)</td>
<td>(.019)</td>
</tr>
<tr>
<td>Poor Democracies</td>
<td>.373</td>
<td>.393</td>
</tr>
<tr>
<td></td>
<td>(.094)</td>
<td>(.081)</td>
</tr>
<tr>
<td>N</td>
<td>1966</td>
<td>5627</td>
</tr>
</tbody>
</table>

Replication of Baum and Lake; the effect of being a democracy on life expectancy and on the percentage enrolled in secondary education (with p-values in parentheses).