Soc504: Mixtures, EM and Missing Data

Brandon Stewart[1]

Princeton

March 27- April 5, 2017

[1] The 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)

Readings

- Monday (Mixture Models)
  - Imai, Kosuke, and Dustin Tingley. "A statistical method for empirical testing of competing theories." *American Journal of Political Science* 56.1 (2012): 218-236.

Readings

- Monday (Mixture Models)
  - Imai, Kosuke, and Dustin Tingley. "A statistical method for empirical testing of competing theories." *American Journal of Political Science* 56.1 (2012): 218-236.
  - Bishop, Christopher. *Pattern Recognition and Machine Learning* (2006). Chapter 9.1-9.2

Readings

- Monday (Mixture Models)
  - Imai, Kosuke, and Dustin Tingley. "A statistical method for empirical testing of competing theories." *American Journal of Political Science* 56.1 (2012): 218-236.
  - Bishop, Christopher. *Pattern Recognition and Machine Learning* (2006). Chapter 9.1-9.2
  - Garip, Filiz. "Discovering diverse mechanisms of migration: The Mexico-US Stream 1970-2000." *Population and Development Review* 38.3 (2012): 393-433. (Optional)

Readings

- Monday (Mixture Models)
  - Imai, Kosuke, and Dustin Tingley. "A statistical method for empirical testing of competing theories." *American Journal of Political Science* 56.1 (2012): 218-236.
  - Bishop, Christopher. *Pattern Recognition and Machine Learning* (2006). Chapter 9.1-9.2
  - Garip, Filiz. "Discovering diverse mechanisms of migration: The Mexico-US Stream 1970-2000." *Population and Development Review* 38.3 (2012): 393-433. (Optional)
- Wednesday (EM)

Readings

- Monday (Mixture Models)
  - Imai, Kosuke, and Dustin Tingley. "A statistical method for empirical testing of competing theories." *American Journal of Political Science* 56.1 (2012): 218-236.
  - Bishop, Christopher. *Pattern Recognition and Machine Learning* (2006). Chapter 9.1-9.2
  - Garip, Filiz. "Discovering diverse mechanisms of migration: The Mexico-US Stream 1970-2000." *Population and Development Review* 38.3 (2012): 393-433. (Optional)
- Wednesday (EM)
  - Bishop, Christopher. *Pattern Recognition and Machine Learning* (2006). Chapter 9 (Optional)

Readings

- Monday (Mixture Models)
  - Imai, Kosuke, and Dustin Tingley. "A statistical method for empirical testing of competing theories." *American Journal of Political Science* 56.1 (2012): 218-236.
  - Bishop, Christopher. *Pattern Recognition and Machine Learning* (2006). Chapter 9.1-9.2
  - Garip, Filiz. "Discovering diverse mechanisms of migration: The Mexico-US Stream 1970-2000." *Population and Development Review* 38.3 (2012): 393-433. (Optional)
- Wednesday (EM)
  - Bishop, Christopher. *Pattern Recognition and Machine Learning* (2006). Chapter 9 (Optional)
- Monday (Missing Data)

Readings

- Monday (Mixture Models)
  - Imai, Kosuke, and Dustin Tingley. "A statistical method for empirical testing of competing theories." *American Journal of Political Science* 56.1 (2012): 218-236.
  - Bishop, Christopher. *Pattern Recognition and Machine Learning* (2006). Chapter 9.1-9.2
  - Garip, Filiz. "Discovering diverse mechanisms of migration: The Mexico-US Stream 1970-2000." *Population and Development Review* 38.3 (2012): 393-433. (Optional)
- Wednesday (EM)
  - Bishop, Christopher. *Pattern Recognition and Machine Learning* (2006). Chapter 9 (Optional)
- Monday (Missing Data)
  - King, Gary; James Honaker; Ann Joseph; Kenneth Scheve. 2001. “Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation,” *American Political Science Review* 95, 1 (March 2001): 49-69.

Readings

- Monday (Mixture Models)
  - Imai, Kosuke, and Dustin Tingley. "A statistical method for empirical testing of competing theories." *American Journal of Political Science* 56.1 (2012): 218-236.
  - Bishop, Christopher. *Pattern Recognition and Machine Learning* (2006). Chapter 9.1-9.2
  - Garip, Filiz. "Discovering diverse mechanisms of migration: The Mexico-US Stream 1970-2000." *Population and Development Review* 38.3 (2012): 393-433. (Optional)
- Wednesday (EM)
  - Bishop, Christopher. *Pattern Recognition and Machine Learning* (2006). Chapter 9 (Optional)
- Monday (Missing Data)
  - King, Gary; James Honaker; Ann Joseph; Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *American Political Science Review* 95, 1 (March 2001): 49-69.
  - James Honaker and Gary King. "What to do about Missing Values in Time Series Cross-Section Data," *American Journal of Political Science* 54, 2 (April, 2010): 561-581 (Optional)

Readings

- Monday (Mixture Models)
  - Imai, Kosuke, and Dustin Tingley. "A statistical method for empirical testing of competing theories." *American Journal of Political Science* 56.1 (2012): 218-236.
  - Bishop, Christopher. *Pattern Recognition and Machine Learning* (2006). Chapter 9.1-9.2
  - Garip, Filiz. "Discovering diverse mechanisms of migration: The Mexico-US Stream 1970-2000." *Population and Development Review* 38.3 (2012): 393-433. (Optional)
- Wednesday (EM)
  - Bishop, Christopher. *Pattern Recognition and Machine Learning* (2006). Chapter 9 (Optional)
- Monday (Missing Data)
  - King, Gary; James Honaker; Ann Joseph; Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *American Political Science Review* 95, 1 (March 2001): 49-69.
  - James Honaker and Gary King. "What to do about Missing Values in Time Series Cross-Section Data," *American Journal of Political Science* 54, 2 (April, 2010): 561-581 (Optional)
- Wednesday (Missing Data)

Readings

- Monday (Mixture Models)
  - Imai, Kosuke, and Dustin Tingley. "A statistical method for empirical testing of competing theories." *American Journal of Political Science* 56.1 (2012): 218-236.
  - Bishop, Christopher. *Pattern Recognition and Machine Learning* (2006). Chapter 9.1-9.2
  - Garip, Filiz. "Discovering diverse mechanisms of migration: The Mexico-US Stream 1970-2000." *Population and Development Review* 38.3 (2012): 393-433. (Optional)
- Wednesday (EM)
  - Bishop, Christopher. *Pattern Recognition and Machine Learning* (2006). Chapter 9 (Optional)
- Monday (Missing Data)
  - King, Gary; James Honaker; Ann Joseph; Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *American Political Science Review* 95, 1 (March 2001): 49-69.
  - James Honaker and Gary King. "What to do about Missing Values in Time Series Cross-Section Data," *American Journal of Political Science* 54, 2 (April, 2010): 561-581 (Optional)
- Wednesday (Missing Data)
  - Blackwell, Matthew, James Honaker, and Gary King. 2014. "A Unified Approach to Measurement Error and Missing Data: Overview, Details and Extensions" *Sociological Methods and Research* (Optional)

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

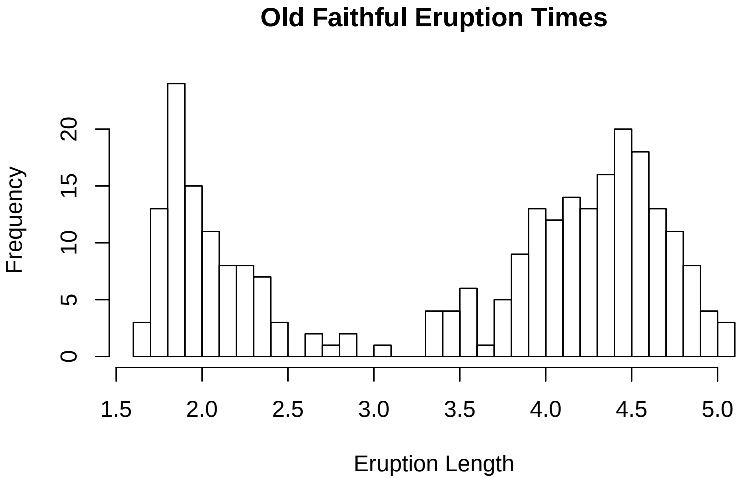
2. Expectation Maximization
   - EM for Probit Regression
   - EM for Gaussian Mixtures
   - EM in General

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

4. Measurement Error

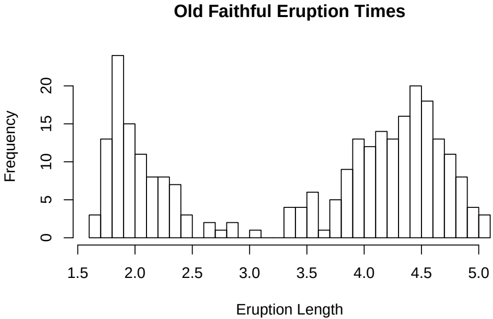
5. Appendix: Additional Details and Examples

Old Faithful



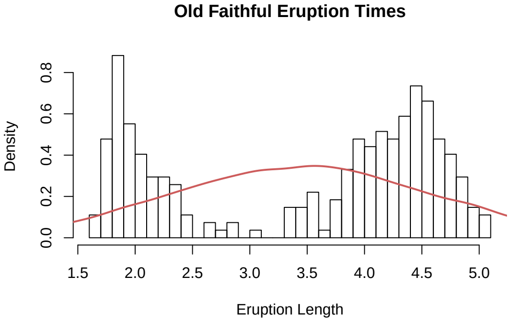
Old Faithful

Old Faithful



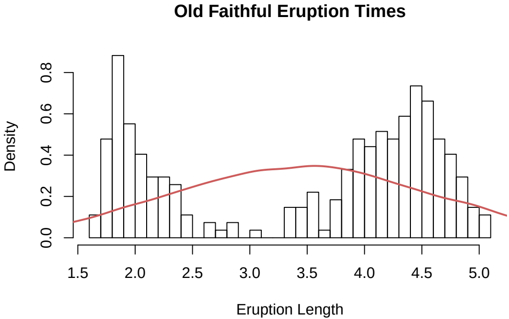
- How do we summarize? No handy distribution

Old Faithful



- 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.

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 $z_i \sim$Bernoulli(π)

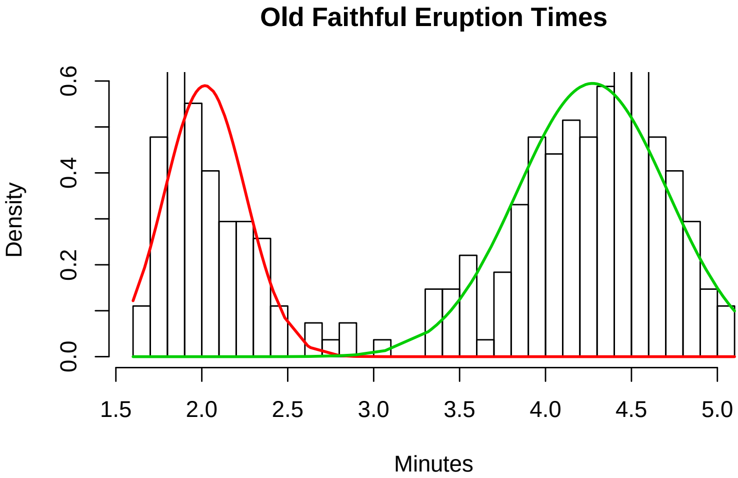
Mixture Models

- 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 $z_i \sim$Bernoulli(π)
- 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 $z_i \sim$Bernoulli(π)
- 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



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.

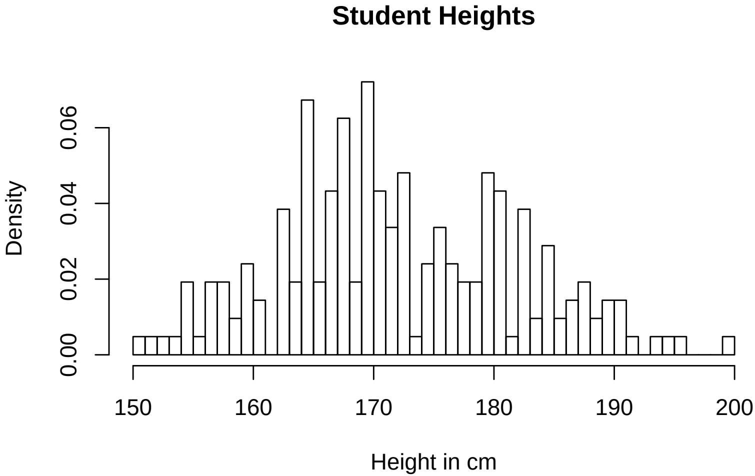
Mixture Models

- 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

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

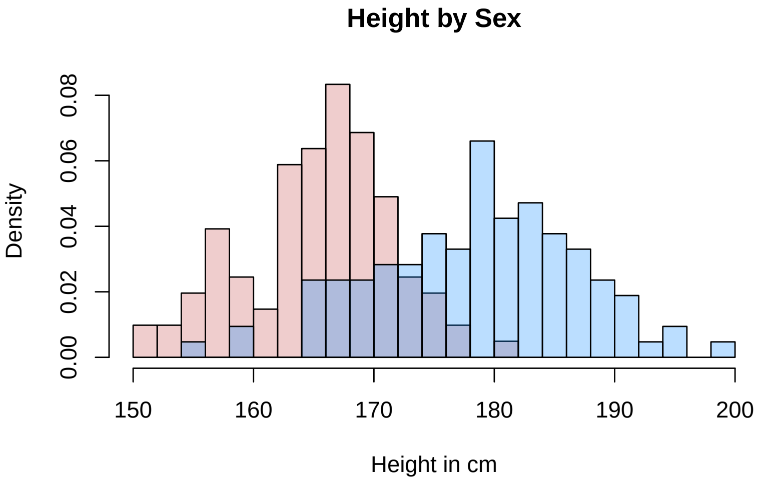
- This is hard to solve due to the summation inside the log
- By introducing the missing variables z, we make it easier to estimate the parameters. This is called data augmentation.
- This problem was easy because the components are well separated.

A Harder Problem



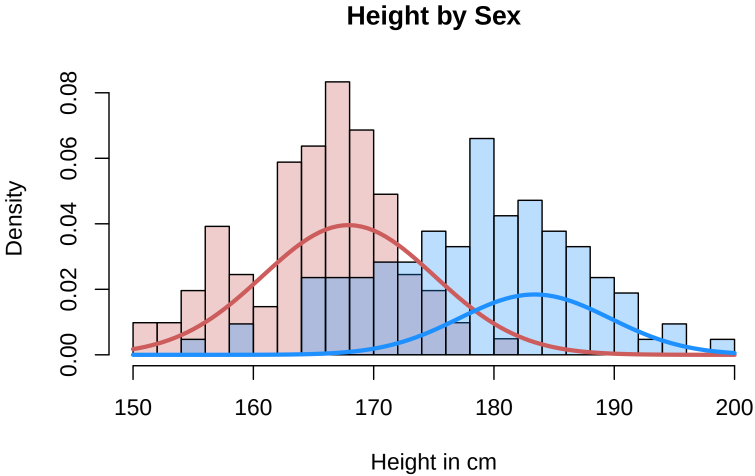
Some distributions have less clear separation

A Harder Problem



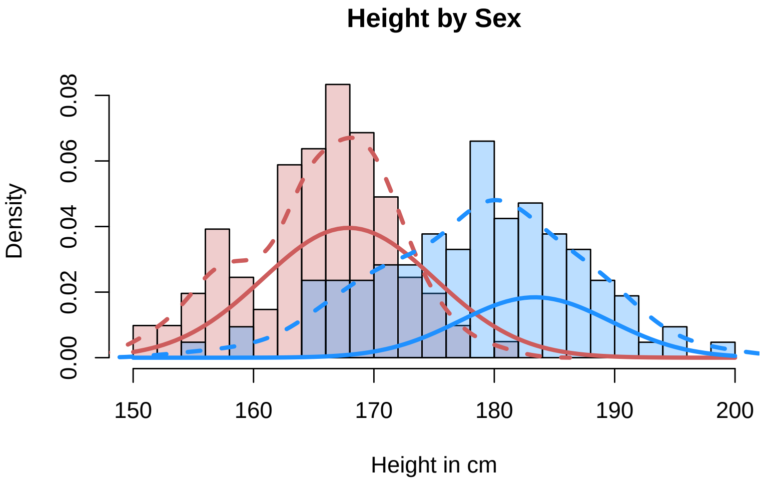
Bimodality here arises due to gender

A Harder Problem



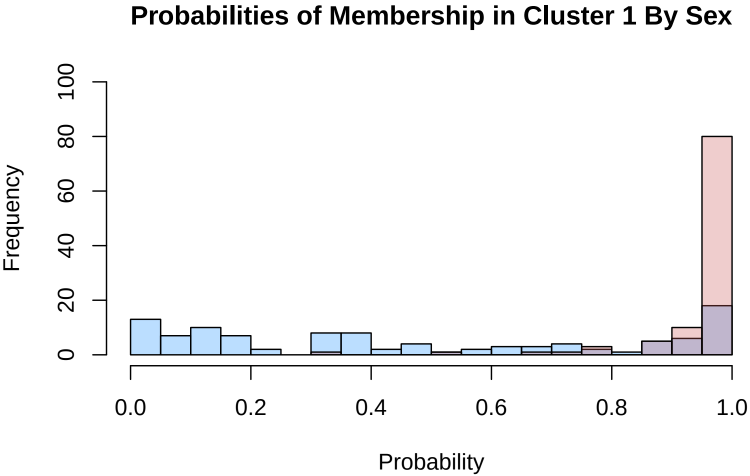
The mixture model *sort of* captures this

A Harder Problem



The true distributions are more peaked with fatter tails

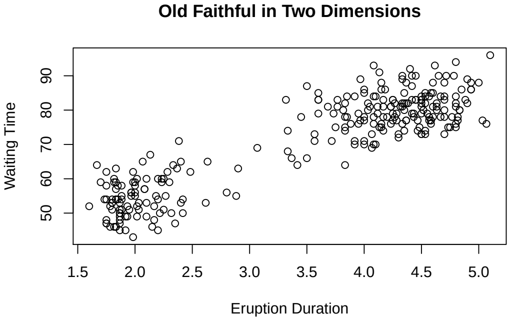
A Harder Problem



One component captures all the women but also many men

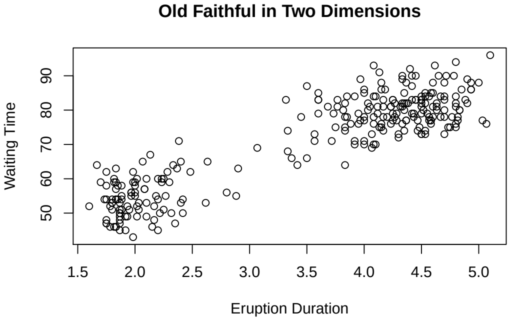
Multiple Dimensions

Multiple Dimensions



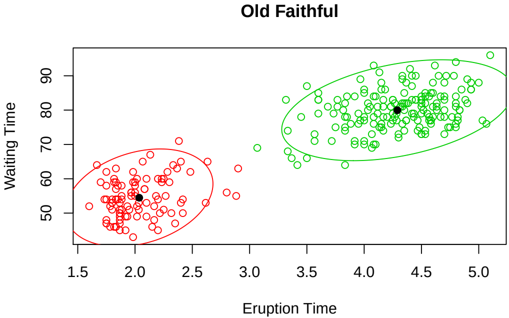
- This strategy also works in more than one dimension

Multiple Dimensions



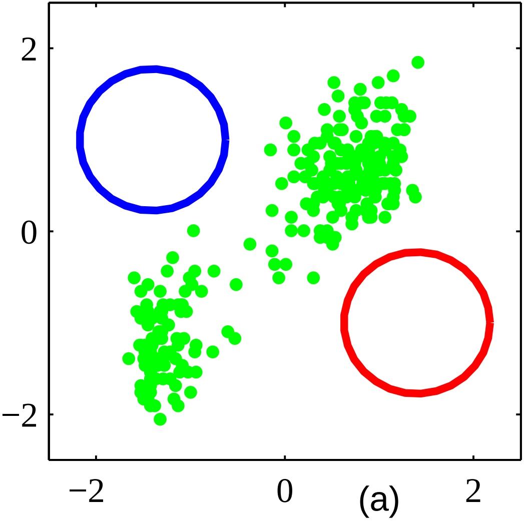
- This strategy also works in more than one dimension
- Now the cluster indicator indexes a multivariate distribution

Multiple Dimensions



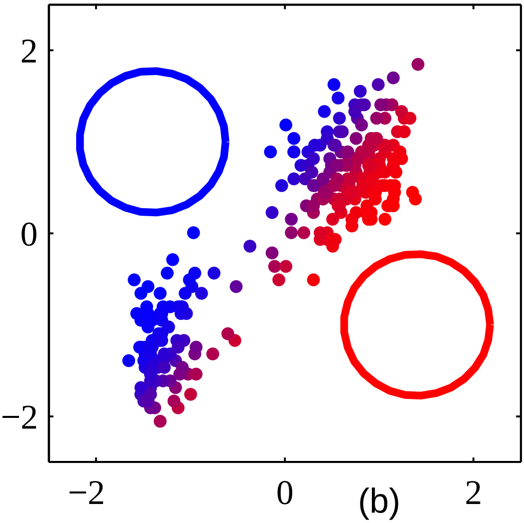
- This strategy also works in more than one dimension
- Now the cluster indicator indexes a multivariate distribution
- This fits the data reasonable well

The Gist of Computation



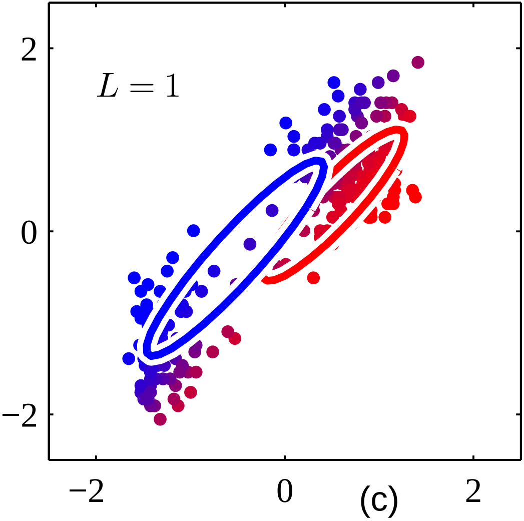
From Bishop (2006) Chapter 9

The Gist of Computation



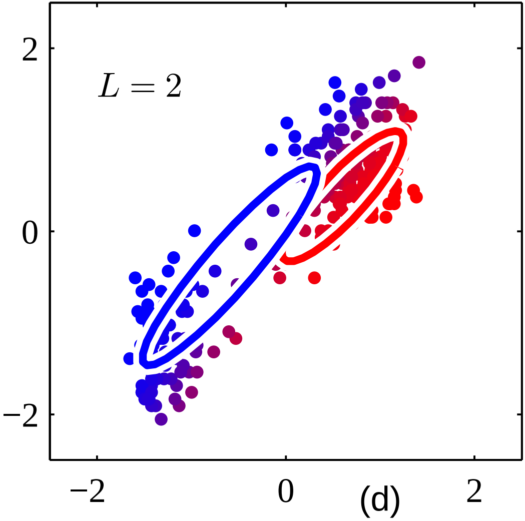
From Bishop (2006) Chapter 9

The Gist of Computation



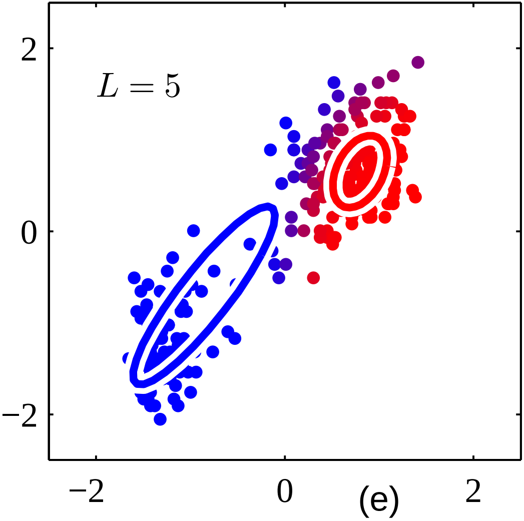
From Bishop (2006) Chapter 9

The Gist of Computation



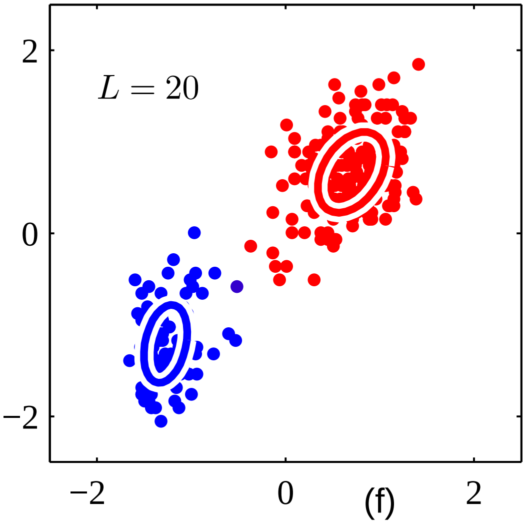
From Bishop (2006) Chapter 9

The Gist of Computation



From Bishop (2006) Chapter 9

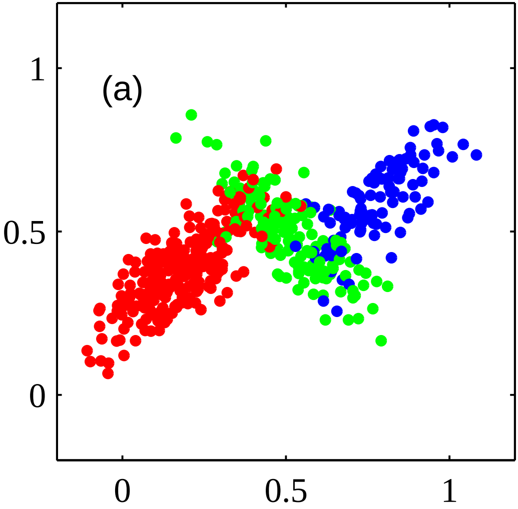
The Gist of Computation



From Bishop (2006) Chapter 9

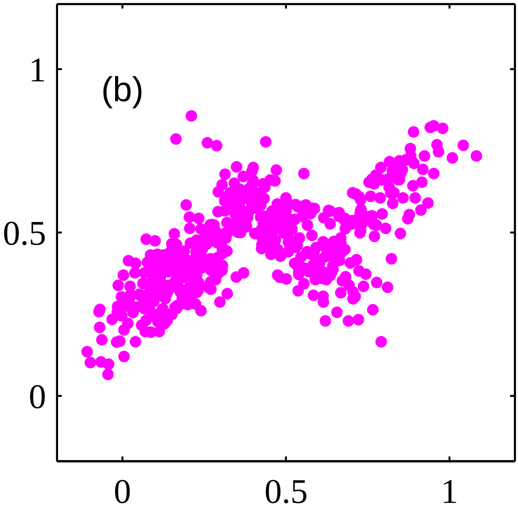
Mixture Models Can Have Many Components

Mixture Models Can Have Many Components



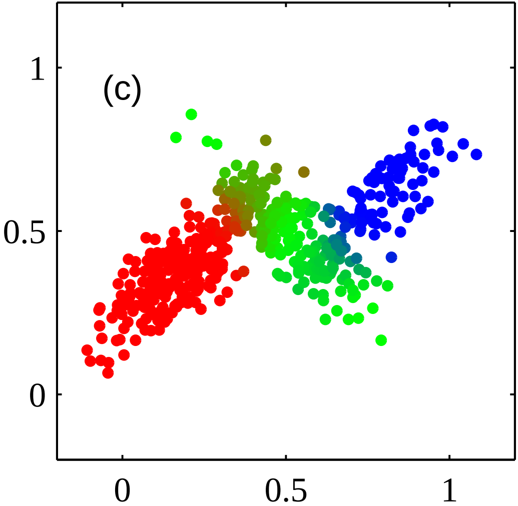
Imagine we draw from data with a 3 component mixture

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

- Garip (2012) “Discovering Diverse Mechanisms of Migration: The Mexico-U.S. Stream 1970-2000”

Two Applications

- Garip (2012) “Discovering Diverse Mechanisms of Migration: The Mexico-U.S. Stream 1970-2000”
- Imai and Tingley (2012) “A Statistical Method for Empirical Testing of Competing Theories”

Two Applications

- Garip (2012) “Discovering Diverse Mechanisms of Migration: The Mexico-U.S. Stream 1970-2000”
- Imai and Tingley (2012) “A Statistical Method for Empirical Testing of Competing Theories”
- Two articles motivated from a common methodological place

Two Applications

- Garip (2012) "Discovering Diverse Mechanisms of Migration: The Mexico-U.S. Stream 1970-2000"
- Imai and Tingley (2012) "A Statistical Method for Empirical Testing of Competing Theories"
- Two articles motivated from a common methodological place
- Both use mixtures in the context of regression

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.

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)

- 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)

- 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)

- 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:

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.
- 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)

- 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

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.
- 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

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.
- 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)

- 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)

- 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.

Details

- Garip (2012) uses an iterative clustering algorithm related to *k*-means which minimizes the sum of distances between data points and a cluster center.

Details

- 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?

Details

- 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.

Details

- 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

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)

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.

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 $\rightsquigarrow$ *k*-means (Banerjee et al 2005, Kulis and Jordan 2012).

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 $\rightsquigarrow$ *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

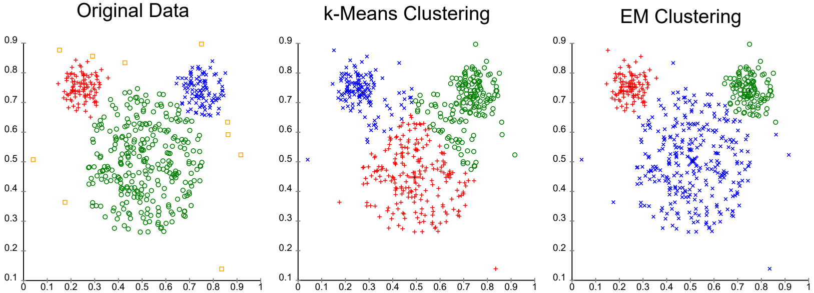
- We started class with the example of a mixture model with Normally distributed components, often called a Gaussian Mixture Model (GMM)
- *k*-means typically minimizes the L_2 (Euclidean distance) which shares the squared-loss objective with the Gaussian distribution.
- We can obtain a correspondence between the two using small-variance asymptotics. As the covariances of the Gaussian go to zero, the EM algorithm for the GMM $\rightsquigarrow$ *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

Connections: *k*-means and Gaussian Mixtures

The biggest impact is that *k*-means strongly prefers equal sized clusters.

Different cluster analysis results on "mouse" data set:



Results

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

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

Next Steps

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

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...

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 π).

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

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 π).
- 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) $\rightsquigarrow$ factor owners 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

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) $\rightsquigarrow$ 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

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:

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

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

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

Overview

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

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:

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

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
  - M-Step: update the model parameters by maximizing the complete data likelihood
- We will step through a few cases to see how this works.

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

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

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.

The Intuition

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 β?

The Intuition

What if we observed Y_i^*?

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

How do we estimate β?

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

The Intuition

What if we observed Y_i^*?

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

How do we estimate β?

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

But oh yeah, we don't know Y_i^*

The Intuition

What if we knew β?

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

The Intuition

What if we knew β?

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

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 β?

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

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

$$\begin{aligned} E(Y_i^*|y_i, X_i, \beta) &= E(X_i\beta + \epsilon_i|y_i, X_i, \beta) \\ &= E(X_i\beta|y_i, X_i, \beta) + E(\epsilon_i|y_i, X_i, \beta) \\ &= X_i\beta + E(\epsilon_i|y_i, X_i, \beta) \end{aligned}$$

The Intuition

What if we knew β?

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

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

$$\begin{aligned} E(Y_i^*|y_i, X_i, \beta) &= E(X_i\beta + \epsilon_i|y_i, X_i, \beta) \\ &= E(X_i\beta|y_i, X_i, \beta) + E(\epsilon_i|y_i, X_i, \beta) \\ &= X_i\beta + E(\epsilon_i|y_i, X_i, \beta) \end{aligned}$$

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 β which we will call β^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 β which we will call β^t.

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

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 β which we will call β^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 β which we will call β^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

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:

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?

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 β^t

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

$$\begin{aligned}
E(Y_i^*|y_i, X_i, \beta^t) &= E(X_i\beta^t + \epsilon_i|y_i, X_i, \beta) \\
&= X_i\beta^t + E(\epsilon_i|y_i, X_i, \beta) \\
&= X_i\beta^t + \left(\frac{-\phi_i(-X_i\beta^t)}{\Phi_i(-X_i\beta^t)}\right)^{(1-y_i)} \left(\frac{\phi_i(-X_i\beta^t)}{(1-\Phi_i(-X_i\beta^t))}\right)^{y_i}
\end{aligned}$$

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

$$\begin{aligned} E(Y_i^*|y_i, X_i, \beta^t) &= E(X_i\beta^t + \epsilon_i|y_i, X_i, \beta) \\ &= X_i\beta^t + E(\epsilon_i|y_i, X_i, \beta) \\ &= X_i\beta^t + \left(\frac{-\phi_i(-X_i\beta^t)}{\Phi_i(-X_i\beta^t)}\right)^{(1-y_i)} \left(\frac{\phi_i(-X_i\beta^t)}{(1-\Phi_i(-X_i\beta^t))}\right)^{y_i} \end{aligned}$$

3. Calculate the estimate for β^{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 β^t
2. Calculate the $E(Y_i^*|y_i, X_i, \beta^t)$

$$\begin{aligned} E(Y_i^*|y_i, X_i, \beta^t) &= E(X_i\beta^t + \epsilon_i|y_i, X_i, \beta) \\ &= X_i\beta^t + E(\epsilon_i|y_i, X_i, \beta) \\ &= X_i\beta^t + \left(\frac{-\phi_i(-X_i\beta^t)}{\Phi_i(-X_i\beta^t)}\right)^{(1-y_i)} \left(\frac{\phi_i(-X_i\beta^t)}{(1-\Phi_i(-X_i\beta^t))}\right)^{y_i} \end{aligned}$$

3. Calculate the estimate for β^{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 θ.

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 θ.
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 θ.
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 θ.
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 $Z^{(t)} = E(Z|\theta^{(t)}, Y)$

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 θ.
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. $\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\ Q(\theta, \theta^{t})$

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 θ.
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. $\theta^{(t+1)} = \underset{\theta}{\text{argmax}}\ Q(\theta, \theta^t)$
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:

$$\boldsymbol{x}_i \quad \sim \quad \text{Distribution(parameters)}$$

Example 2: Mixtures

Single distribution data generating process:

$$\boldsymbol{x}_i \sim \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

Example 2: Mixtures

Single distribution data generating process:

$$\boldsymbol{x}_i \quad \sim \quad \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

$$\boldsymbol{z}_i | \boldsymbol{\pi} \quad \sim \quad \text{Multinomial}(1, \boldsymbol{\pi})$$

Example 2: Mixtures

Single distribution data generating process:

$$\boldsymbol{x}_i \quad \sim \quad \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

$$\begin{aligned} \boldsymbol{z}_i|\boldsymbol{\pi} \quad &\sim \quad \text{Multinomial}(1, \boldsymbol{\pi}) \\ \boldsymbol{x}_i|z_{ik} = 1 \quad &\sim \quad \text{Distribution}(\text{parameters}_k) \end{aligned}$$

Example 2: Mixtures

Single distribution data generating process:

$$\boldsymbol{x}_i \sim \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

$$\begin{aligned} \boldsymbol{z}_i | \boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\ \boldsymbol{x}_i | z_{ik} = 1 &\sim \text{Distribution}(\text{parameters}_k) \end{aligned}$$

In words:

Example 2: Mixtures

Single distribution data generating process:

$$\boldsymbol{x}_i \sim \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

$$\begin{aligned} \boldsymbol{z}_i | \boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\ \boldsymbol{x}_i | z_{ik} = 1 &\sim \text{Distribution}(\text{parameters}_k) \end{aligned}$$

In words:

- Draw a cluster label

Example 2: Mixtures

Single distribution data generating process:

$$\boldsymbol{x}_i \sim \text{Distribution(parameters)}$$

Mixture of distribution data generating process:

$$\begin{aligned} \boldsymbol{z}_i | \boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\ \boldsymbol{x}_i | z_{ik} = 1 &\sim \text{Distribution}(\text{parameters}_k) \end{aligned}$$

In words:

- Draw a cluster label
- Given distribution, draw realization

Gaussian Mixture

Gaussian Mixture

$$\boldsymbol{z}_i | \boldsymbol{\pi} \quad \sim \quad \text{Multinomial}(1, \boldsymbol{\pi})$$

Gaussian Mixture

$$\begin{aligned} \boldsymbol{z}_i | \boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi}) \\ \boldsymbol{x}_i | z_{ik} = 1, \boldsymbol{\mu}_k, \Sigma_k &\sim \text{Normal}(\boldsymbol{\mu}_k, \Sigma_k) \end{aligned}$$

Gaussian Mixture

$$\begin{aligned}\boldsymbol{z}_i|\boldsymbol{\pi} &\sim \text{Multinomial}(1, \boldsymbol{\pi})\\ \boldsymbol{x}_i|z_{ik} = 1, \boldsymbol{\mu}_k, \Sigma_k &\sim \text{Normal}(\boldsymbol{\mu}_k, \Sigma_k)\end{aligned}$$

This leads to the likelihood:

$$\begin{aligned}p(x) &= \sum_z p(z)p(x|z)\\ &= \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)\end{aligned}$$

Algorithm for the Gaussian Mixture

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

Algorithm for the Gaussian Mixture

1) Initialize parameters $\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute 'responsibilities' $p(\boldsymbol{z}_i|\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$

Algorithm for the Gaussian Mixture

1) Initialize parameters $\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute 'responsibilities' $p(\boldsymbol{z}_i|\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$

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

Algorithm for the Gaussian Mixture

1) Initialize parameters $\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute 'responsibilities' $p(\boldsymbol{z}_i | \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$

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

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

Algorithm for the Gaussian Mixture

1) Initialize parameters $\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t$, $\boldsymbol{\pi}^t$
2) Expectation step: compute 'responsibilities' $p(\boldsymbol{z}_i | \boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$

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

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

$$\begin{aligned} \mathrm{E}_z[\log p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\pi})] &= \mathrm{E}_z \left[\log \left(\prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(x_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}} \right) \right] \\ &= \mathrm{E}_z \left[\sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left[\log \pi_k \mathcal{N}(x_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right] \right] \end{aligned}$$

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

Algorithm for the Gaussian Mixture

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

2) Expectation step: compute 'responsibilities' $p(\boldsymbol{z}_i|\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t, \boldsymbol{\pi}^t, \boldsymbol{X}) \rightsquigarrow \boldsymbol{r}_i^t$

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

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

$$\begin{aligned} \mathrm{E}_z[\log p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\pi})] &= \mathrm{E}_z\left[\log\left(\prod_{i=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(x_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}\right)\right] \\ &= \mathrm{E}_z\left[\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left[\log \pi_k \mathcal{N}(x_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right]\right] \end{aligned}$$

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

4) Assess change in the log-likelihood

Algorithm for the Gaussian Mixture

3) M-Step:

Algorithm for the Gaussian Mixture

3) M-Step:

$$\text{E}[\log \text{Complete data} | \boldsymbol{\theta}, \boldsymbol{\pi}] \quad = \quad \sum_{i=1}^{N} \sum_{k=1}^{K} E[z_{ik}] \log \left(\pi_k \mathcal{N}(x_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)$$

Algorithm for the Gaussian Mixture

3) M-Step:

$$\text{E}[\log \text{Complete data}|\boldsymbol{\theta}, \boldsymbol{\pi}] = \sum_{i=1}^{N} \sum_{k=1}^{K} E[z_{ik}] \log \left(\pi_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)$$

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

Algorithm for the Gaussian Mixture

3) M-Step:

$$\text{E}[\log \text{Complete data}|\boldsymbol{\theta}, \boldsymbol{\pi}] = \sum_{i=1}^{N} \sum_{k=1}^{K} E[z_{ik}] \log\left(\pi_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)$$

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

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

Algorithm for the Gaussian Mixture

3) M-Step:

$$\text{E[log Complete data}|\boldsymbol{\theta}, \boldsymbol{\pi}] = \sum_{i=1}^{N}\sum_{k=1}^{K} E[z_{ik}] \log\left(\pi_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)$$

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

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

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

Algorithm for the Gaussian Mixture

3) M-Step:

$$\text{E[log Complete data}|\boldsymbol{\theta}, \boldsymbol{\pi}] = \sum_{i=1}^{N}\sum_{k=1}^{K} E[z_{ik}] \log\left(\pi_k \mathcal{N}(x_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)$$

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

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

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

$$\Sigma_k^{t+1} = \frac{1}{\sum_{i=1}^N r_{ik}^t} \sum_{i=1}^{N} r_{ik}(x_i - \boldsymbol{\mu}_k^{t+1})(x_i - \boldsymbol{\mu}_k^{t+1})^T \quad (3)$$

The EM Algorithm in Words

The EM Algorithm in Words

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

Under the maximum likelihood framework we want to find θ 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 θ describes the joint distribution of x and z, as $p(x, z|\theta)$.

Under the maximum likelihood framework we want to find θ which maximizes:

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

We assume that maximizing the likelihood isn't easy but we can find θ 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 θ describes the joint distribution of x and z, as $p(x, z|\theta)$.

Under the maximum likelihood framework we want to find θ which maximizes:

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

We assume that maximizing the likelihood isn't easy but we can find θ 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 $\boldsymbol{\theta}^t$

The EM Algorithm in Math

1) Initialize parameters $\boldsymbol{\theta}^t$
2) E step: Using the current value of the parameter θ compute the expected value of the log-likelihood with respect to the conditional distribution of $Z|X$

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

The EM Algorithm in Math

1) Initialize parameters $\boldsymbol{\theta}^t$
2) E step: Using the current value of the parameter θ compute the expected value of the log-likelihood with respect to the conditional distribution of $Z|X$

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

3) M step: maximize the Q function:

The EM Algorithm in Math

1) Initialize parameters $\boldsymbol{\theta}^t$
2) E step: Using the current value of the parameter θ compute the expected value of the log-likelihood with respect to the conditional distribution of $Z|X$

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

3) M step: maximize the Q function:

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} Q(\theta|\theta^t)$$

The EM Algorithm in Math

1) Initialize parameters $\boldsymbol{\theta}^t$

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

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

3) M step: maximize the Q function:

$$\theta^{(t+1)} = \underset{\theta}{\mathrm{argmax}} Q(\theta|\theta^t) \tag{5}$$

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

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.

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.

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.

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

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.

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

	Independence		
Attendance	Yes	No	DK
Yes	1439	78	159
No	16	16	32
DK	144	54	136

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.

	Independence	
Attendance	Yes	No
Yes	θ_{11}	θ_{12}
No	θ_{21}	θ_{22}

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:

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:

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.

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:

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 $\hat{\theta}_{11} = \frac{1439}{1549+525} = .6938$.

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.

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:

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.

Imputation

Here's the data again, with the proportion of observed data filled in.

	Independence		
Attendance	Yes	No	DK
Yes	1439 (.928)	78 (.050)	159
No	16 (.010)	16 (.010)	32
DK	144	54	136

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.

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.

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

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.

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

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

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.

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

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.

Imputation: An Updated Sense of the Proportions?

	Independence	
Attendance	Yes	No
Yes	1439 + 150.87 + 142.42	78 + 8.12 + 44.81
	.896	.066
No	16 + 16 + 1.58	16 + 16 + 9.19
	.017	.020

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

Imputation: An Updated Sense of the Proportions?

	Independence	
Attendance	Yes	No
Yes	$1439 + 150.87 + 142.42$	$78 + 8.12 + 44.81$
	.896	.066
No	$16 + 16 + 1.58$	$16 + 16 + 9.19$
	.017	.020

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 θ. What would be a suitable next step?

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.

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

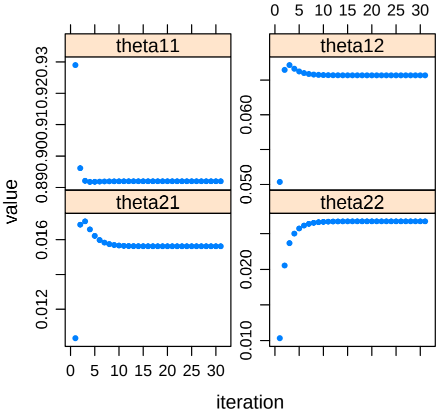
We can now use our updated (and, in fact, improved) estimate of the population proportions in order to re-impute the missing data using the same approach as before.

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 θ evolve through the iterations:



A Final Estimate

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

A Final Estimate

After running the algorithm for 30 iterations, the final estimate for θ_{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 θ_{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 θ_{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

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 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

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

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

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

Missingness Notation

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

Missingness Notation

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

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

Missingness Notation

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

D_{mis} = *missing* elements in D (in Red)
D_{obs} = observed elements in D

Missingness Notation

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

D_{mis} = *missing* elements in D (in Red)
D_{obs} = observed elements in D

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

What Can Be Learned With Minimal Assumptions

What Can Be Learned With Minimal Assumptions

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

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 $Y_i^* = Y_i M_i + (\text{NA})(1 - M_i)$ which simply suggests that the Y_i is stable (e.g. regardless of how the question is asked or who responded).

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 $Y_i^* = Y_i M_i + (\text{NA})(1 - M_i)$ 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.
- This only works with bounded support and becomes much harder with missingness on many variables

Possible Further Assumptions

Assumption	Acronym	You can predict M with:

Possible Further Assumptions

Assumption	Acronym	You can predict M with:

Possible Further Assumptions

Assumption	Acronym	You can predict M with:
Missing Completely At Random	MCAR	—

Possible Further Assumptions

Assumption	Acronym	You can predict M with:
Missing Completely At Random	MCAR	—
Missing At Random	MAR	D_{obs}

Possible Further Assumptions

		You can predict
Assumption	Acronym	M with:
Missing Completely At Random	MCAR	—
Missing At Random	MAR	D_{obs}
Nonignorable	NI	D_{obs} & D_{mis}

Possible Further Assumptions

Assumption	Acronym	You can predict M with:
Missing Completely At Random	MCAR	—
Missing At Random	MAR	D_{obs}
Nonignorable	NI	D_{obs} & D_{mis}

- Reasons for the odd terminology are historical.

Missingness Assumptions, again

Missingness Assumptions, again

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

Missingness Assumptions, again

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

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

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)

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)

$$\mathbb{P}(M|D) \equiv \mathbb{P}(M|D_{obs}, D_{mis}) = \mathbb{P}(M|D_{obs})$$

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)

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)

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)

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

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 $> \$100K$ <u>and</u> you can't predict high income with other measured variables

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 $> \$100K$ <u>and</u> 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

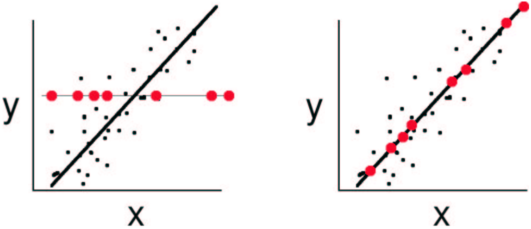
5. Mean Imputation
   severely distorts distribution, pulls correlations to zero
6. Best guess imputation or logical rules
   depends on guesser/logic
7. Indicator for Continuous Variable (aka dummy for missingness)
   biased on other predictors, can include interactions between indicators and other 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

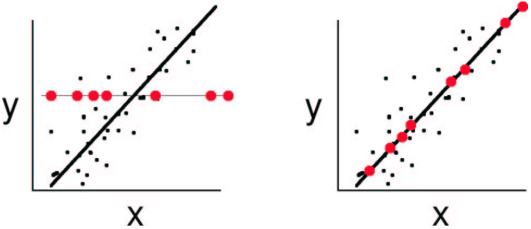
Random Imputation for one variable

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

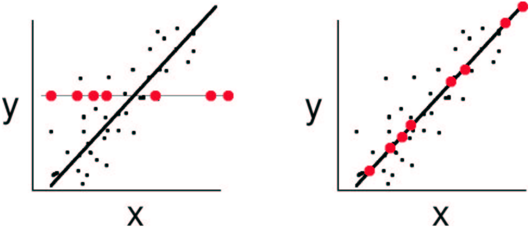
Random Imputation for one variable

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


y
x
y
x



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

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

Application-specific Methods for Missing Data

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

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.

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

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

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

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?

How to create application-specific methods?

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

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

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

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

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

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 and M 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

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

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

$$\begin{aligned} &= P(D_{obs}|\theta)P(M|D_{obs}, \gamma), \\ &\propto P(D_{obs}|\theta) \end{aligned}$$

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

$$\begin{aligned} &= P(D_{obs}|\theta)P(M|D_{obs}, \gamma), \\ &\propto P(D_{obs}|\theta) \end{aligned}$$

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

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

$$\begin{aligned} &= P(D_{obs}|\theta)P(M|D_{obs}, \gamma), \\ &\propto P(D_{obs}|\theta) \end{aligned}$$

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

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

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

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

$$\begin{aligned} &= P(D_{obs}|\theta)P(M|D_{obs}, \gamma), \\ &\propto P(D_{obs}|\theta) \end{aligned}$$

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

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

$$\begin{aligned} &= P(D_{obs}|\theta)P(M|D_{obs}, \gamma), \\ &\propto P(D_{obs}|\theta) \end{aligned}$$

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

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

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

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

Multiple Imputation

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

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:

$$\text{SE}(q)^2 = \text{mean}(\text{SE}_j^2) + \text{variance}(q_j)\,(1 + 1/m)$$

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

$$\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j$$

Standard error: use Rubin's Rule:

$$\begin{aligned} \mathrm{SE}(q)^2 &= \mathrm{mean}(\mathrm{SE}_j^2) + \mathrm{variance}(q_j)\,(1 + 1/m) \\ &= \mathrm{within} + \mathrm{between} \end{aligned}$$

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:

$$\begin{aligned} \mathrm{SE}(q)^2 &= \mathrm{mean}(\mathrm{SE}_j^2) + \mathrm{variance}(q_j)\,(1 + 1/m) \\ &= \mathrm{within} + \mathrm{between} \end{aligned}$$

Last piece vanishes as m increases

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

$$\bar{q} = \frac{1}{m} \sum_{j=1}^{m} q_j$$

Standard error: use Rubin's Rule:

$$\begin{aligned} \mathrm{SE}(q)^2 &= \mathrm{mean}(\mathrm{SE}_j^2) + \mathrm{variance}(q_j)\,(1 + 1/m) \\ &= \mathrm{within} + \mathrm{between} \end{aligned}$$

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

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):

$$L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma)$$

A General Model for Imputations

1. If data were complete, we could use (it's deceptively simple):

$$L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \qquad \text{(a multivariate normal)}$$

A General Model for Imputations

1. If data were complete, we could use (it's deceptively simple):

$$L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \qquad \text{(a multivariate normal)}$$

2. With missing data, this becomes:

A General Model for Imputations

1. If data were complete, we could use (it's deceptively simple):

$$L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \qquad \text{(a multivariate normal)}$$

2. With missing data, this becomes:

$$L(\mu, \Sigma | D_{obs}) \propto \prod_{i=1}^{n} \int N(D_i | \mu, \Sigma) dD_{mis}$$

A General Model for Imputations

1. If data were complete, we could use (it's deceptively simple):

$$L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \qquad \text{(a multivariate normal)}$$

2. With missing data, this becomes:

$$\begin{aligned} L(\mu, \Sigma | D_{obs}) &\propto \prod_{i=1}^{n} \int N(D_i | \mu, \Sigma) dD_{mis} \\ &= \prod_{i=1}^{n} N(D_{i,obs} | \mu_{obs}, \Sigma_{obs}) \end{aligned}$$

A General Model for Imputations

1. If data were complete, we could use (it's deceptively simple):

$$L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \qquad \text{(a multivariate normal)}$$

2. With missing data, this becomes:

$$\begin{aligned} L(\mu, \Sigma | D_{obs}) &\propto \prod_{i=1}^{n} \int N(D_i | \mu, \Sigma) dD_{mis} \\ &= \prod_{i=1}^{n} N(D_{i,obs} | \mu_{obs}, \Sigma_{obs}) \end{aligned}$$

since marginals of MVN's are normal.

A General Model for Imputations

1. If data were complete, we could use (it's deceptively simple):

$$L(\mu, \Sigma | D) \propto \prod_{i=1}^{n} N(D_i | \mu, \Sigma) \qquad \text{(a multivariate normal)}$$

2. With missing data, this becomes:

$$\begin{aligned} L(\mu, \Sigma | D_{obs}) &\propto \prod_{i=1}^{n} \int N(D_i | \mu, \Sigma) dD_{mis} \\ &= \prod_{i=1}^{n} N(D_{i,obs} | \mu_{obs}, \Sigma_{obs}) \end{aligned}$$

   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 (μ, Σ).

A General Model for Imputations

1. If data were complete, we could use (it's deceptively simple):

$$L(\mu, \Sigma|D) \propto \prod_{i=1}^{n} N(D_i|\mu, \Sigma) \qquad \text{(a multivariate normal)}$$

2. With missing data, this becomes:

$$\begin{aligned} L(\mu, \Sigma|D_{obs}) &\propto \prod_{i=1}^{n} \int N(D_i|\mu, \Sigma) dD_{mis} \\ &= \prod_{i=1}^{n} N(D_{i,obs}|\mu_{obs}, \Sigma_{obs}) \end{aligned}$$

   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 (μ, Σ).
4. Difficult computationally: $D_{i,obs}$ has different elements observed for each i and so each observation is informative about different pieces of (μ, Σ).

5. Difficult Statistically: number of parameters increases quickly in the number of variables (p, columns of D):

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):

$$\begin{aligned} \text{parameters} &= \text{parameters}(\mu) + \text{parameters}(\Sigma) \\ &= p + p(p+1)/2 \end{aligned}$$

5. Difficult Statistically: number of parameters increases quickly in the number of variables (p, columns of D):

$$\begin{aligned}\text{parameters} &= \text{parameters}(\mu) + \text{parameters}(\Sigma)\\ &= p + p(p+1)/2 = p(p+3)/2.\end{aligned}$$

5. Difficult Statistically: number of parameters increases quickly in the number of variables (p, columns of D):

$$\begin{aligned}\text{parameters} &= \text{parameters}(\mu) + \text{parameters}(\Sigma)\\ &= p + p(p+1)/2 = p(p+3)/2.\end{aligned}$$

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):

$$\begin{aligned}\text{parameters} &= \text{parameters}(\mu) + \text{parameters}(\Sigma) \\ &= p + p(p+1)/2 = p(p+3)/2.\end{aligned}$$

   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, columns of D):

$$\begin{aligned} \text{parameters} &= \text{parameters}(\mu) + \text{parameters}(\Sigma) \\ &= p + p(p+1)/2 = p(p+3)/2. \end{aligned}$$

   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

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:

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:

$$\begin{aligned} D &\sim N(D|\mu, \Sigma) \\ &= N\left[\begin{pmatrix} Y \\ X \end{pmatrix} \,\middle|\, \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix}\right] \end{aligned}$$

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:

$$\begin{aligned} D &\sim N(D|\mu, \Sigma) \\ &= N\left[\begin{pmatrix} Y \\ X \end{pmatrix} \,\middle|\, \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix}\right] \end{aligned}$$

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:

$$\begin{aligned} D &\sim N(D|\mu, \Sigma) \\ &= N\left[\begin{pmatrix} Y \\ X \end{pmatrix} \;\middle|\; \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix}\right] \end{aligned}$$

4. Conditionals of bivariate normals are normal:

$$Y|X \sim N\left(y|E(Y|X), V(Y|X)\right)$$

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:

$$\begin{aligned} D &\sim N(D|\mu, \Sigma) \\ &= N\left[\begin{pmatrix} Y \\ X \end{pmatrix} \,\middle|\, \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix}\right] \end{aligned}$$

4. Conditionals of bivariate normals are normal:

$$Y|X \sim N\left(y|E(Y|X), V(Y|X)\right)$$

$$E(Y|X) = \mu_y + \beta(X - \mu_x) \quad \text{(a regression of } Y \text{ on all other } X\text{'s!)}$$

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:

$$\begin{aligned} D &\sim N(D|\mu, \Sigma) \\ &= N\left[\begin{pmatrix} Y \\ X \end{pmatrix} \,\middle|\, \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix}\right] \end{aligned}$$

4. Conditionals of bivariate normals are normal:

$$Y|X \sim N\left(y|E(Y|X), V(Y|X)\right)$$

$$\begin{aligned} E(Y|X) &= \mu_y + \beta(X - \mu_x) \quad \text{(a regression of } Y \text{ on all other } X\text{'s!)} \\ \beta &= \sigma_{xy}/\sigma_x \end{aligned}$$

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:

$$\begin{aligned} D &\sim N(D|\mu, \Sigma) \\ &= N\left[\begin{pmatrix} Y \\ X \end{pmatrix} \middle| \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y & \sigma_{xy} \\ \sigma_{xy} & \sigma_x \end{pmatrix}\right] \end{aligned}$$

4. Conditionals of bivariate normals are normal:

$$Y|X \sim N\left(y|E(Y|X), V(Y|X)\right)$$

$$\begin{aligned} E(Y|X) &= \mu_y + \beta(X - \mu_x) \quad \text{(a regression of } Y \text{ on all other } X\text{'s!)} \\ \beta &= \sigma_{xy}/\sigma_x \\ V(Y|X) &= \sigma_y - \sigma_{xy}^2/\sigma_x \end{aligned}$$

5. To create imputations:

5. To create imputations:
   (a) Estimate the posterior density of μ and Σ

5. To create imputations:
    (a) Estimate the posterior density of μ and Σ
        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 μ and Σ
        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

5. To create imputations:
    (a) Estimate the posterior density of μ and Σ
        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 μ and Σ from their posterior density

5. To create imputations:
    (a) Estimate the posterior density of μ and Σ
        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 μ and Σ from their posterior density
    (c) Compute simulations of $E(Y|X)$ and $V(Y|X)$ deterministically

5. To create imputations:
    (a) Estimate the posterior density of μ and Σ
        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 μ and Σ 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 μ and Σ
      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 μ and Σ 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

EMB: EM With Bootstrap

EMB: EM With Bootstrap

- 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 β and Σ in each (for estimation uncertainty)

EMB: EM With Bootstrap

- Randomly draw n obs (with replacement) from the data
- Use EM to estimate β and Σ 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 β and Σ in each (for estimation uncertainty)
- Impute D_{mis} from each from the model (for fundamental uncertainty)
- Lightning fast; works with very large data sets

EMB: EM With Bootstrap

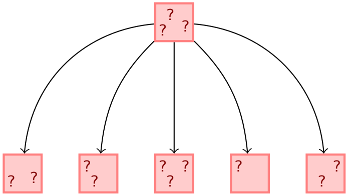
- Randomly draw n obs (with replacement) from the data
- Use EM to estimate β and Σ in each (for estimation uncertainty)
- Impute D_{mis} from each from the model (for fundamental uncertainty)
- Lightning fast; works with very large data sets
- Basis for Amelia II

Multiple Imputation: Amelia Style

Multiple Imputation: Amelia Style

incomplete data

Multiple Imputation: Amelia Style

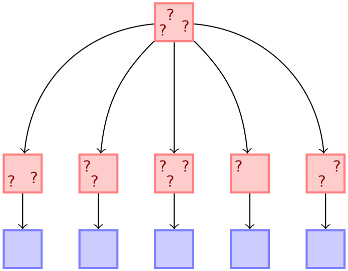


incomplete data

bootstrap

bootstrapped data

Multiple Imputation: Amelia Style



incomplete data

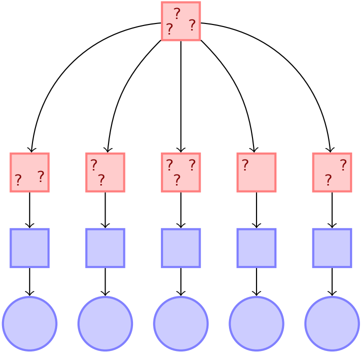
bootstrap

bootstrapped data

EM

imputed datasets

Multiple Imputation: Amelia Style



incomplete data

bootstrap

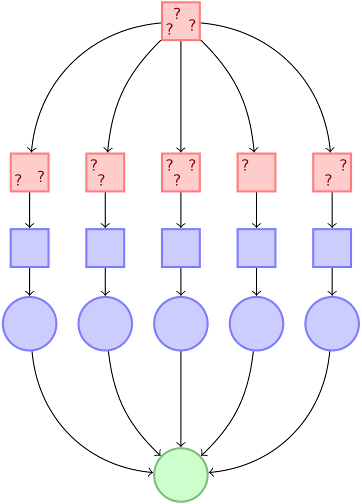
bootstrapped data

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

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!

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.

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 $\leadsto$ 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)

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: $\sqrt{\text{counts}}$, $\ln(p/(1-p))$, $\ln(\text{money})$, etc.

What Can Go Wrong and What to Do

- Inference is learning about facts we don't have with facts we have; we <u>assume</u> 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: $\sqrt{\text{counts}}$, $\ln(p/(1-p))$, $\ln(\text{money})$, etc.
- Code ordinal variables as close to interval as possible.

What Can Go Wrong and What to Do (continued)

What Can Go Wrong and What to Do (continued)

- 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)

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")

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

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)?

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?

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 $\mathbb{AMELIA}$ $\mathbb{II}$ Do?

The $\mathbb{AMELIA}$ $\mathbb{II}$ 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:

$$(y, X) \sim MVN(\mu, \Sigma).$$

Let's define (y,X) = D to simplify the notation.

Once again this gives us two full conditionals for our unknowns (μ, Σ, D_{mis}):

1. $p(D_{Mis}|\mu, \Sigma, D_{Obs})$
2. $L(\mu, \Sigma|D_{Obs}, D_{Mis})$

What Does $\mathbb{AMELIA}$ $\mathbb{II}$ Do?

The $\mathbb{AMELIA}$ $\mathbb{II}$ 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 (μ, Σ, D_{mis}):

1. $p(D_{Mis}|\mu, \Sigma, D_{Obs})$
2. $L(\mu, \Sigma|D_{Obs}, D_{Mis})$

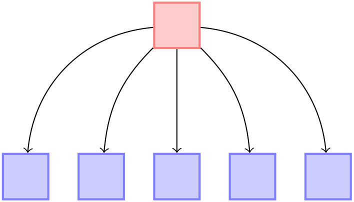
The EM algorithm in this case involves selecting an initial value for (μ, Σ), using that value to impute the missing data, and then re-estimating (μ, Σ) based on the (now-complete) data.

The Multiple Imputation Scheme

The Multiple Imputation Scheme

incomplete data

The Multiple Imputation Scheme

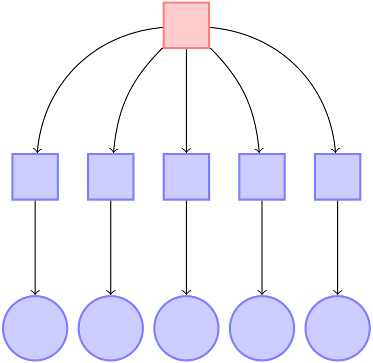


incomplete data

imputation

imputed datasets

The Multiple Imputation Scheme



incomplete data

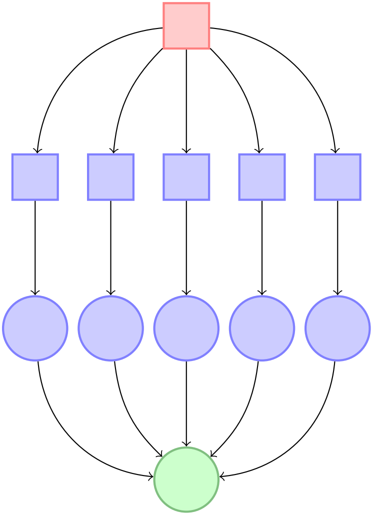
imputation

imputed datasets

analysis

separate results

The Multiple Imputation Scheme



incomplete data

imputation

imputed datasets

analysis

separate results

combination

final results

Multiple Imputation

Multiple Imputation

REGRESSION

To preserve the relationships in the data.

Multiple Imputation

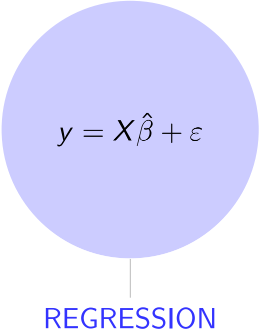
REGRESSION

To preserve the relationships in the data.

SIMULATION

To reflect the uncertainty of our imputation.

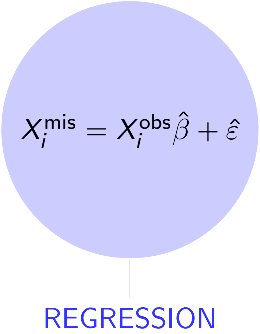
How to Impute



$$y = X\hat{\beta} + \varepsilon$$

REGRESSION

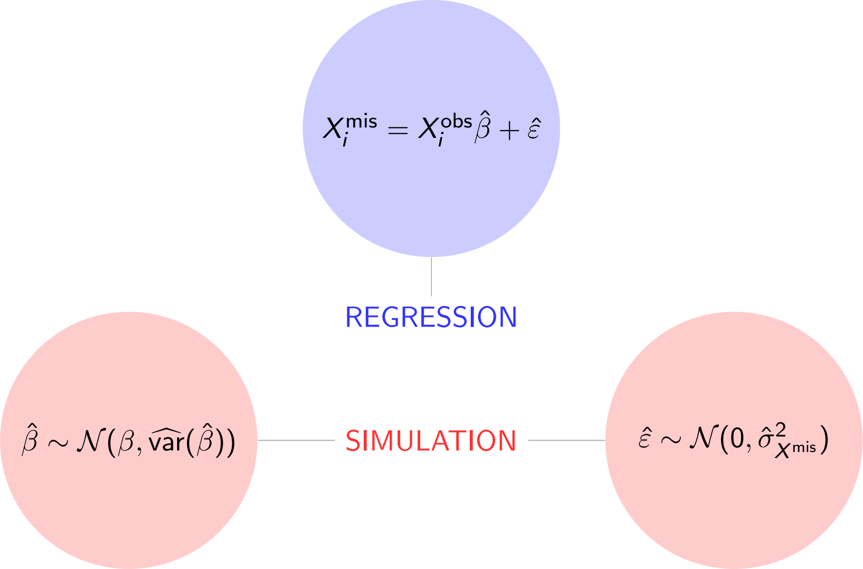
How to Impute



$$X_i^{\text{mis}} = X_i^{\text{obs}}\hat{\beta} + \hat{\varepsilon}$$

REGRESSION

How to Impute



$$X_i^{\text{mis}} = X_i^{\text{obs}}\hat{\beta} + \hat{\varepsilon}$$

REGRESSION

$$\hat{\beta} \sim \mathcal{N}(\beta, \widehat{\text{var}}(\hat{\beta}))$$

SIMULATION

$$\hat{\varepsilon} \sim \mathcal{N}(0, \hat{\sigma}^2_{X^{\text{mis}}})$$

Patterns of Missingness

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	NA	NA	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	NA	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

$$infl = \beta_0 + \beta_1 \cdot GDP + \beta_2 \cdot population + \varepsilon$$

Patterns of Missingness

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	NA	NA	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	NA	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

$$trade = \beta_0 + \beta_1 \cdot GDP + \beta_2 \cdot population + \varepsilon$$

Patterns of Missingness

	year	country	GDP	infl	trade	population
1	1972	Burkina Faso	377	-2.92	29.69	5848380
2	1973	Burkina Faso	376	7.60	31.31	5958700
3	1974	Burkina Faso	393	NA	NA	6075700
4	1975	Burkina Faso	416	18.76	40.11	6202000
5	1976	Burkina Faso	435	NA	37.76	6341030
6	1977	Burkina Faso	448	29.99	41.11	6486870

$$\mathit{infl} = \beta_0 + \beta_1 \cdot \mathit{GDP} + \beta_2 \cdot \mathit{trade} + \beta_3 \cdot \mathit{population} + \varepsilon$$

Any β is just (μ, Σ)

- 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 (μ, Σ) to estimate X^{mis}.
2. Use those estimates of X^{mis} and X^{obs} to get a new estimate of (μ, Σ).
3. Iterate until convergence.

(μ_t, Σ_t)

The EM algorithm

Turn a hard problem into a repeated easy problem.

1. Use current estimates of (μ, Σ) to estimate X^{mis}.
2. Use those estimates of X^{mis} and X^{obs} to get a new estimate of (μ, Σ).
3. Iterate until convergence.

$$\begin{tikzcd}
 & \mathbb{E}[X^{\text{mis}}_{t+1}] \\
(\mu_t, \Sigma_t) \arrow[ur] &
\end{tikzcd}$$

The EM algorithm

Turn a hard problem into a repeated easy problem.

1. Use current estimates of (μ, Σ) to estimate X^{mis}.
2. Use those estimates of X^{mis} and X^{obs} to get a new estimate of (μ, Σ).
3. Iterate until convergence.

$$\begin{tikzcd}
 & \mathbb{E}[X_{t+1}^{\text{mis}}] \arrow[d] \\
(\mu_t, \Sigma_t) \arrow[ur] & (\mu_{t+1}, \Sigma_{t+1}) \\
 & X^{\text{obs}} \arrow[u]
\end{tikzcd}$$

The EM algorithm

Turn a hard problem into a repeated easy problem.

1. Use current estimates of (μ, Σ) to estimate X^{mis}.
2. Use those estimates of X^{mis} and X^{obs} to get a new estimate of (μ, Σ).
3. Iterate until convergence.

$$\begin{tikzcd}
 & \mathbb{E}[X_{t+1}^{\text{mis}}] \arrow[d] & \mathbb{E}[X_{t+2}^{\text{mis}}] \\
(\mu_t, \Sigma_t) \arrow[ur] & (\mu_{t+1}, \Sigma_{t+1}) \arrow[ur] & \\
 & X^{\text{obs}} \arrow[u] &
\end{tikzcd}$$

The EM algorithm

Turn a hard problem into a repeated easy problem.

1. Use current estimates of (μ, Σ) to estimate X^{mis}.
2. Use those estimates of X^{mis} and X^{obs} to get a new estimate of (μ, Σ).
3. Iterate until convergence.

$$\begin{tikzcd}
 & \mathbb{E}[X^{\text{mis}}_{t+1}] \arrow[d] & \mathbb{E}[X^{\text{mis}}_{t+2}] \arrow[d] \\
(\mu_t, \Sigma_t) \arrow[ur] & (\mu_{t+1}, \Sigma_{t+1}) \arrow[ur] & (\mu_{t+2}, \Sigma_{t+2}) \\
 & X^{\text{obs}} \arrow[u] & X^{\text{obs}} \arrow[u]
\end{tikzcd}$$

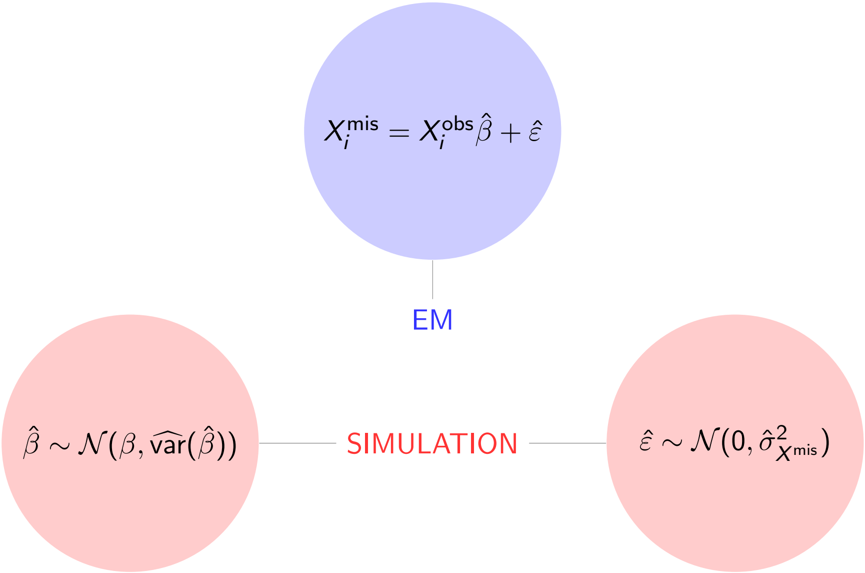
Simulation

$$\underbrace{X_i^{\text{mis}}}_{\text{missing values in row i}} = \underbrace{X_i^{\text{obs}}\beta}_{\text{observed values in row i}} + \underbrace{\varepsilon}_{\mathcal{N}(0,\sigma^2)}$$

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



$X_i^{\text{mis}} = X_i^{\text{obs}}\hat{\beta} + \hat{\varepsilon}$

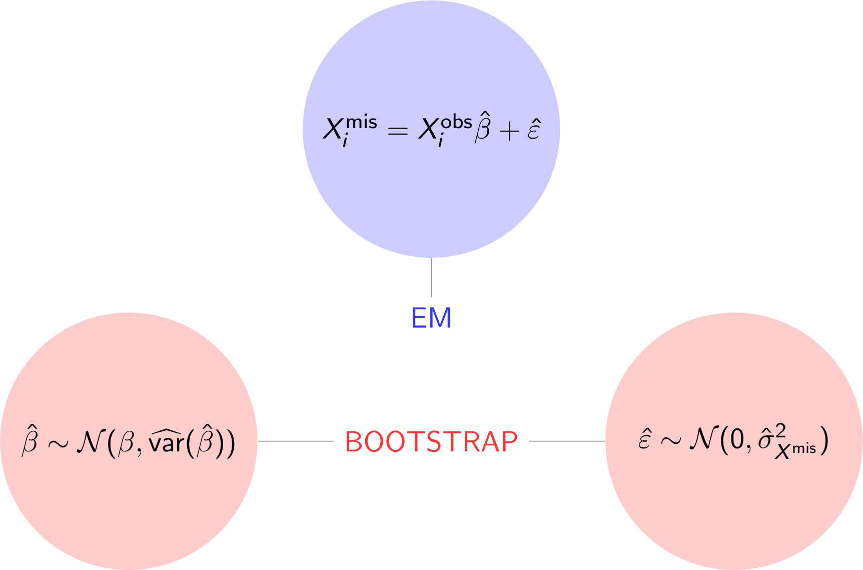
EM

$\hat{\beta} \sim \mathcal{N}(\beta, \widehat{\text{var}}(\hat{\beta}))$

SIMULATION

$\hat{\varepsilon} \sim \mathcal{N}(0, \hat{\sigma}^2_{X^{\text{mis}}})$

How to Impute



$$X_i^{\text{mis}} = X_i^{\text{obs}}\hat{\beta} + \hat{\varepsilon}$$

EM

$$\hat{\beta} \sim \mathcal{N}(\beta, \widehat{\text{var}}(\hat{\beta}))$$

BOOTSTRAP

$$\hat{\varepsilon} \sim \mathcal{N}(0, \hat{\sigma}^2_{X^{\text{mis}}})$$

How to Impute

$$\underbrace{X_i^{\text{mis}}}_{\substack{\text{missing values}\\ \text{in row i}}} = \underbrace{X_i^{\text{obs}}}_{\substack{\text{observed values}\\ \text{in row i}}} \beta + \underbrace{\varepsilon}_{\mathcal{N}(0,\sigma^2)}$$

How to Impute

$$\underbrace{X_i^{\text{mis}}}_{\text{missing values in row i}} = \boxed{\underbrace{X_i^{\text{obs}}\beta}_{\text{observed values in row i}}} + \underbrace{\varepsilon}_{\mathcal{N}(0,\sigma^2)}$$

REGRESSION

How to Impute

$$\underbrace{X_i^{\text{mis}}}_{\text{missing values in row i}} = \underbrace{X_i^{\text{obs}}\beta}_{\text{observed values in row i}} + \underbrace{\varepsilon}_{\mathcal{N}(0,\sigma^2)}$$

REGRESSION SIMULATION

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.

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

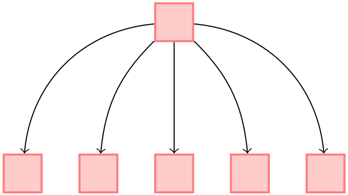
- We will impute a missing value by drawing from a Normal distribution centered around what its predicted by a regression of that variable on the available data in that observation.
- A hard part is the regression, as we have to run a regression for every missing value in every pattern of missingness.
- This could be a lot of regressions, depending on the data.

The $\mathbb{A}$melia Scheme

The $\mathbb{A}$melia Scheme

incomplete data

The 𝔸melia Scheme

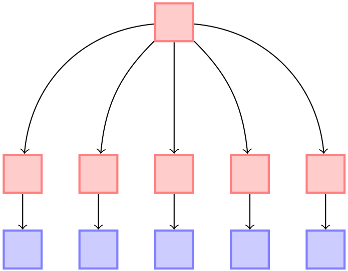


incomplete data

bootstrap

bootstrapped data

The 𝔸melia Scheme



incomplete data

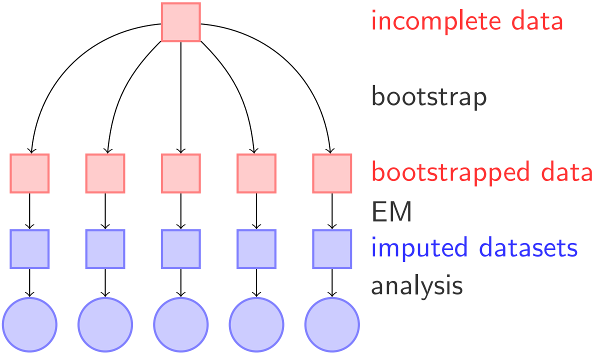
bootstrap

bootstrapped data

EM

imputed datasets

The 𝔸melia Scheme



incomplete data

bootstrap

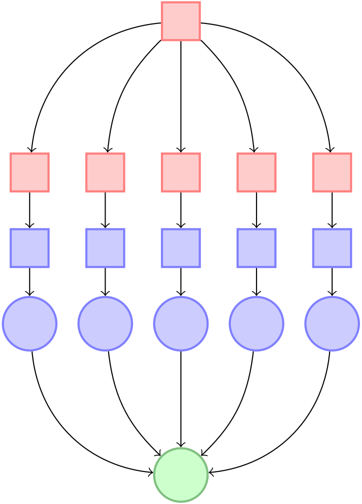
bootstrapped data

EM

imputed datasets

analysis

The 𝔸melia Scheme



incomplete data

bootstrap

bootstrapped data

EM

imputed datasets

analysis

combination

final results

The $\mathbb{A}$melia approach

1. Draw a sample of size n with replacement, X^*.

The $\mathbb{A}$melia 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 $\mathbb{A}$melia 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 $\mathbb{A}$melia 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

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:

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

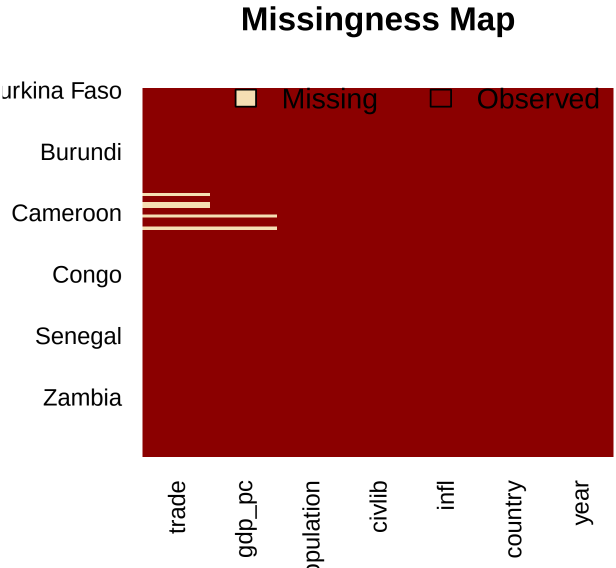
Assumptions and Estimators for Missing Data

- 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

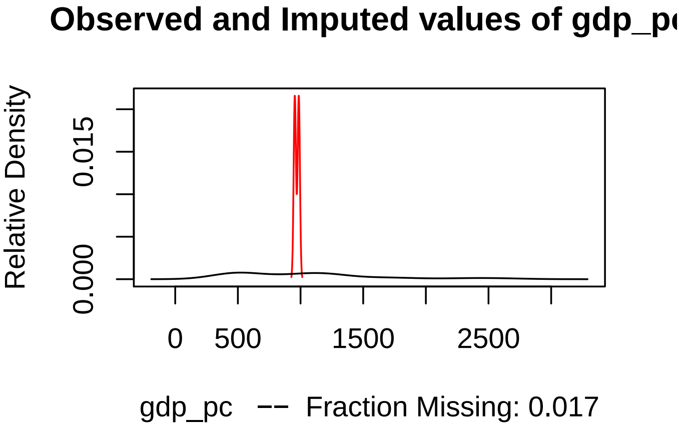
Assumptions and Estimators for Missing Data

- 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

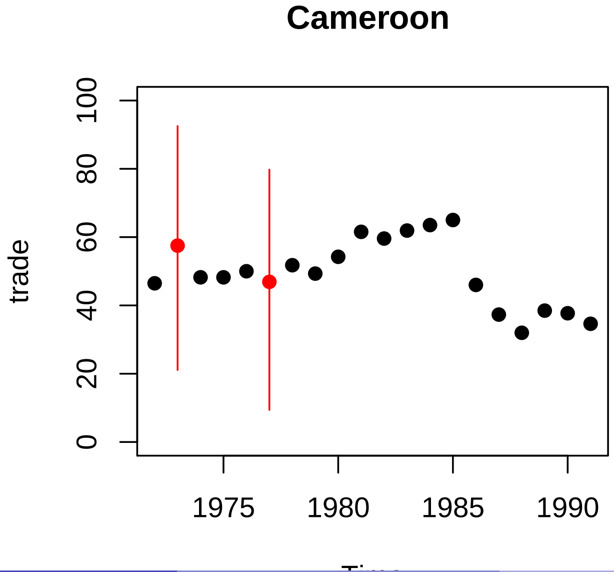


Example Amelia Diagnostics

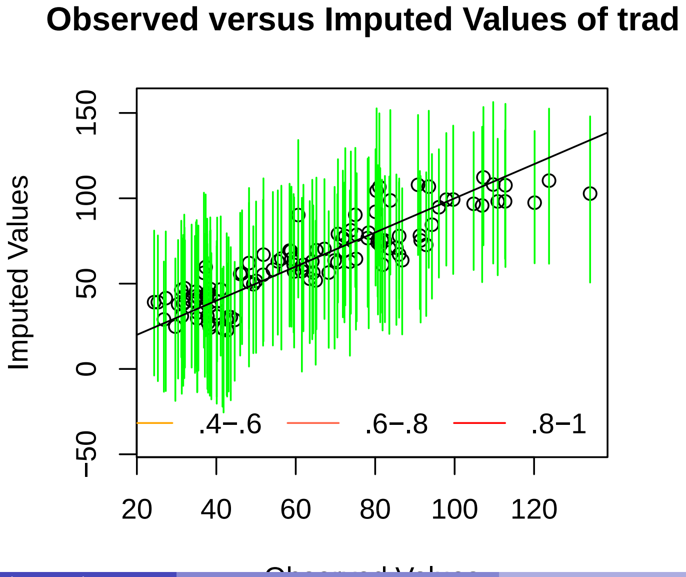


Observed and Imputed values of trade

Example Amelia Diagnostics



Example Amelia Diagnostics



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

Measurement Error or, How Amelia Solves All Your Problems

Blackwell, Matthew, James Honaker, and Gary King. “Multiple Overimputation: A Unified Approach to Measurement Error and Missing Data.” *Sociological Methods and Research* 2015.

Three New Things

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.

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	5	6	6.23	5.92
2	LIBERIA	NA	3	NA	NA
3	SIERRA LEONE	3	3	6.60	NA
4	GHANA	9	6	6.86	12.68
5	TOGO	NA	5	6.27	17.34
6	CAMEROON	6	5	6.93	15.47
7	NIGERIA	5	7	6.88	17.46
8	GABON	6	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	5	6	6.23	5.92
2	LIBERIA	NA	3	NA	NA
3	SIERRA LEONE	3	3	6.60	NA
4	GHANA	9	6	6.86	12.68
5	TOGO	NA	5	6.27	17.34
6	CAMEROON	6	5	6.93	15.47
7	NIGERIA	≈5	7	6.88	17.46
8	GABON	6	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	≈5	6	6.23	5.92
2	LIBERIA	NA	3	NA	NA
3	SIERRA LEONE	≈3	3	6.60	NA
4	GHANA	≈9	6	6.86	12.68
5	TOGO	NA	5	6.27	17.34
6	CAMEROON	≈6	5	6.93	15.47
7	NIGERIA	≈5	7	6.88	17.46
8	GABON	≈6	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	≈5	≈6	6.23	5.92
2	LIBERIA	NA	≈3	NA	NA
3	SIERRA LEONE	≈3	≈3	6.60	NA
4	GHANA	≈9	≈6	6.86	12.68
5	TOGO	NA	≈6	6.27	17.34
6	CAMEROON	≈6	≈5	6.93	15.47
7	NIGERIA	≈5	≈7	6.88	17.46
8	GABON	≈6	≈8	8.19	16.97

One Solution:

	country	GDP	infl	trade	population
1	Ghana	377	-2.92	29.69	5848380
2	Ivory Coast	376	7.60	31.31	5958700
3	Kenya	393	8.72	35.22	6075700
4	Nigeria	416	18.76	40.11	6202000
5	Uganda	435	-8.40	37.76	6341030
6	Burkina Faso	448	29.99	41.11	6486870

One Solution: Change research agendas

	country	~~GDP~~	infl	trade	population
1	Ghana	~~377~~	-2.92	29.69	5848380
2	Ivory Coast	~~376~~	7.60	31.31	5958700
3	Kenya	~~393~~	8.72	35.22	6075700
4	Nigeria	~~416~~	18.76	40.11	6202000
5	Uganda	~~435~~	-8.40	37.76	6341030
6	Burkina Faso	~~448~~	29.99	41.11	6486870

One Solution: Change research agendas

	country	infl	trade	population
1	Ghana	-2.92	29.69	5848380
2	Ivory Coast	7.60	31.31	5958700
3	Kenya	8.72	35.22	6075700
4	Nigeria	18.76	40.11	6202000
5	Uganda	-8.40	37.76	6341030
6	Burkina Faso	29.99	41.11	6486870

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

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 heteroskadastic errors.

The current approaches in the literature

Most existing approaches are application-specific.
model dependent.
difficult to implement.
inapplicable with multiple variables.
invalid with heteroskadastic 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

observed

$$x_i = x_i^* + u_i$$

A Brief Review of Measurement Error

$$\overset{\substack{\text{observed}\\ \searrow}}{x_i} = \overset{\substack{\text{latent}\\ \downarrow}}{x_i^*} + u_i$$

A Brief Review of Measurement Error

$$\overset{\substack{\text{observed}\\ \searrow}}{x_i} = \overset{\substack{\text{latent}\\ \downarrow}}{x_i^*} + \overset{\substack{\text{measurement}\\ \text{error}\\ \swarrow}}{u_i}$$

A Brief Review of Measurement Error

$$\overset{\overset{\text{observed}}{\searrow}}{x_i} \;=\; \overset{\overset{\text{latent}}{\downarrow}}{x_i^*} \;+\; \overset{\overset{\substack{\text{measurement}\\ \text{error}}}{\swarrow}}{u_i}$$

$$u_i|x_i^* \;\sim\; \mathcal{N}(0,\sigma_u^2)$$

A Brief Review of Measurement Error

$$\overset{\substack{\text{observed}\\ \downarrow}}{x_i} \;=\; \overset{\substack{\text{latent}\\ \downarrow}}{x_i^*} \;+\; \overset{\substack{\text{measurement}\\ \text{error}\\ \downarrow}}{u_i}$$

$$u_i|x_i^* \;\sim\; \mathcal{N}(0,\sigma_u^2)$$

unbiased
independent

A Brief Review of Measurement Error

$$\overset{\text{observed}}{x_i} = \overset{\text{latent}}{x_i^*} + \overset{\text{measurement error}}{u_i}$$

$$\underset{\text{unbiased independent}}{u_i|x_i^* \sim \mathcal{N}(0,}\underset{\text{measurement error variance}}{\sigma_u^2})$$

Want to run:

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:

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:

$$y_i = \beta_0 + \beta_1 x_i^* + \beta_2 w_i^* + \beta_3 z_i^* + \epsilon_i$$

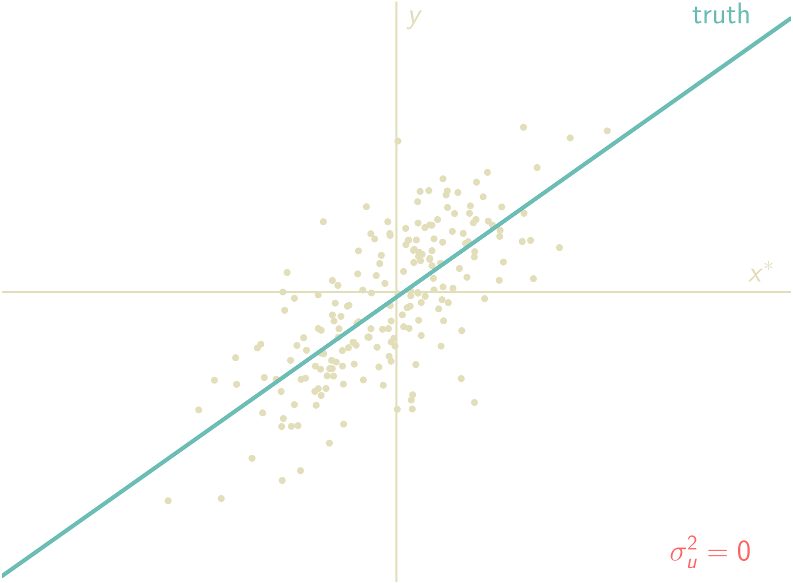
Can only run:

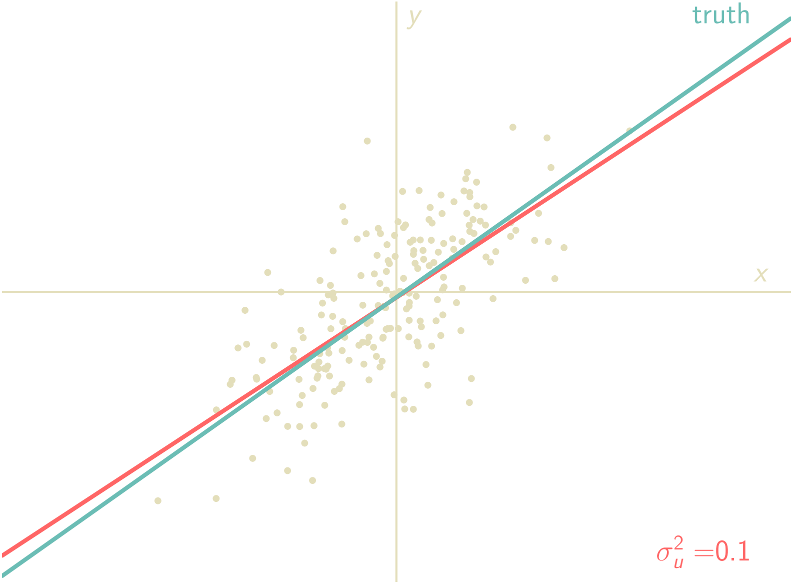
$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 w_i + \alpha_3 z_i + \nu_i$$

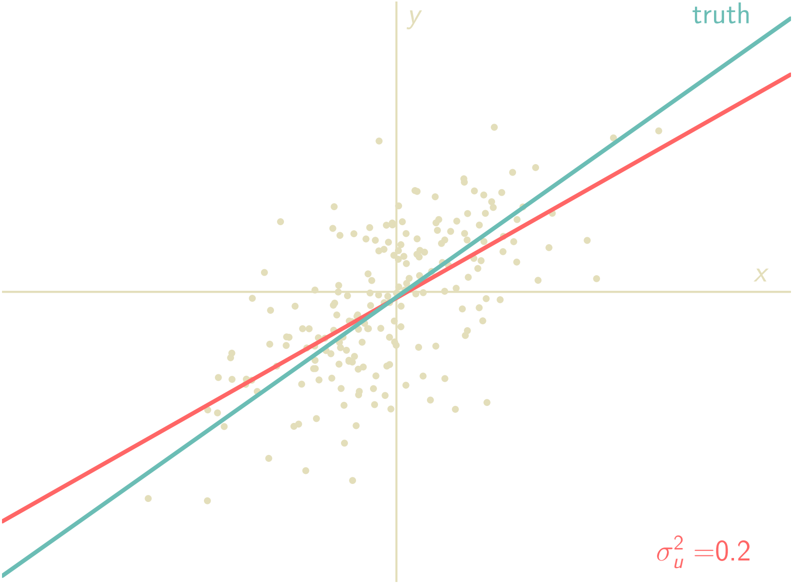
Leads to:

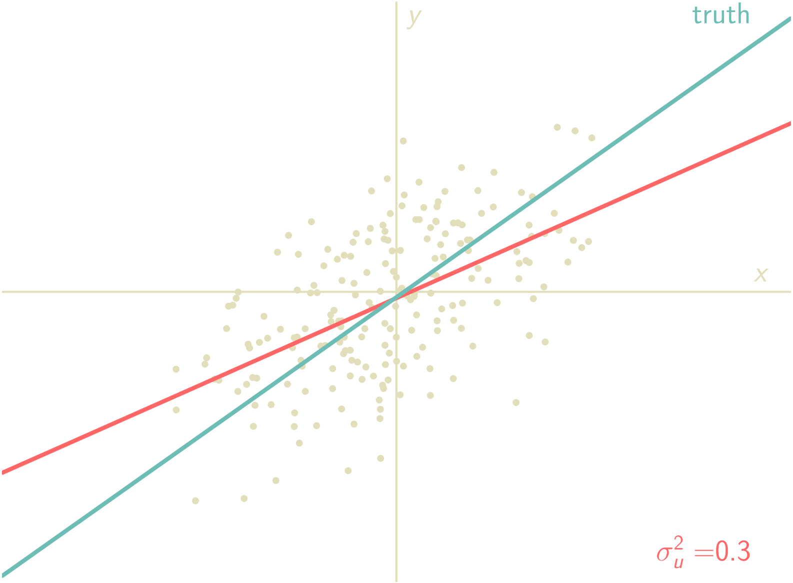
UNKNOWN

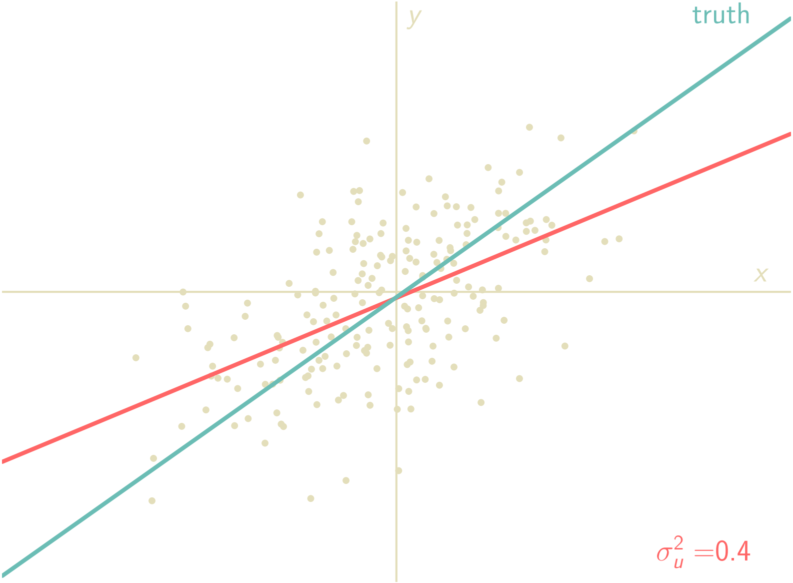
(No guarantees with more mismeasured variables)

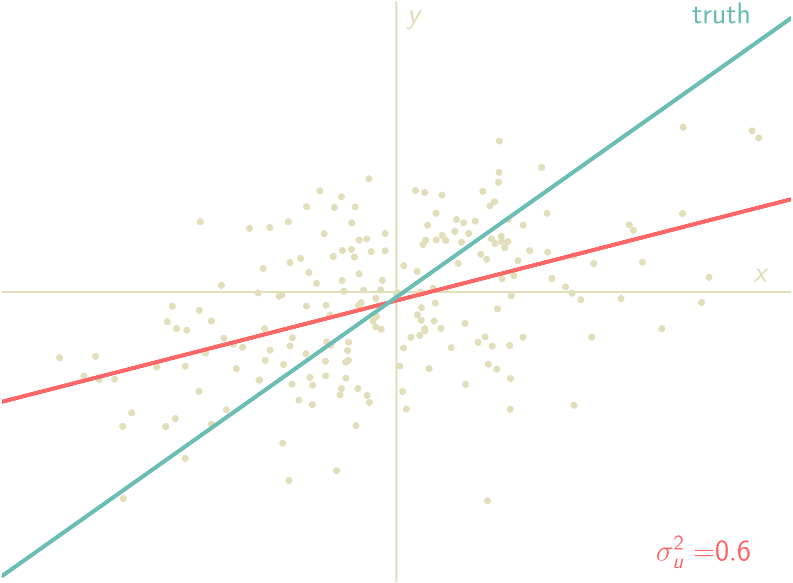

y
truth
x^*
$\sigma_u^2 = 0$

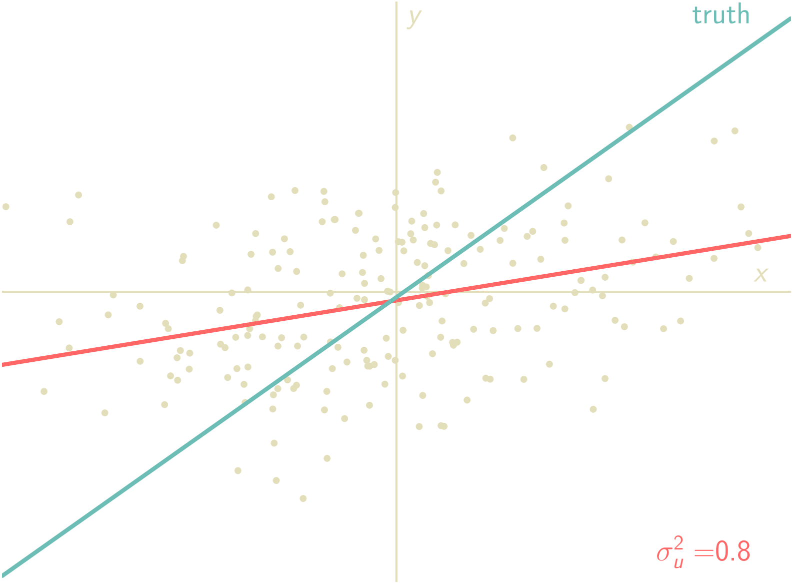

y
truth
x
$\sigma_u^2 = 0.1$

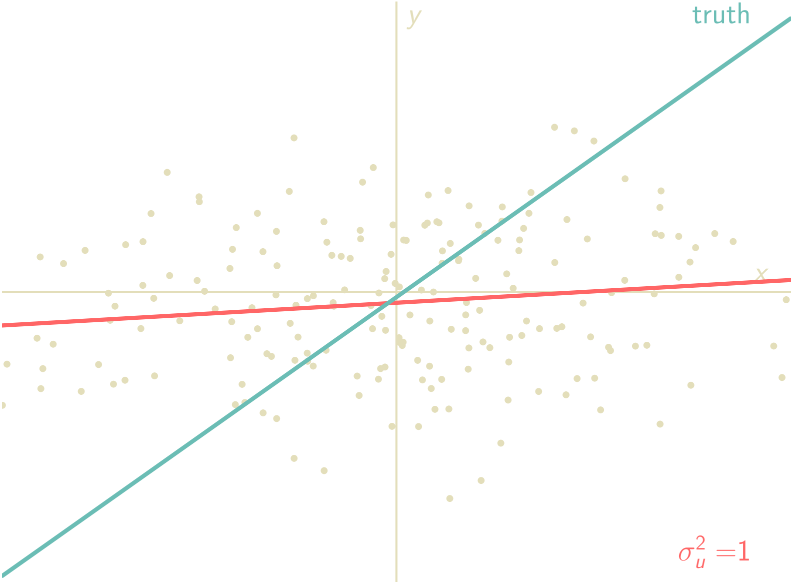

y
truth
x
$\sigma_u^2 = 0.2$

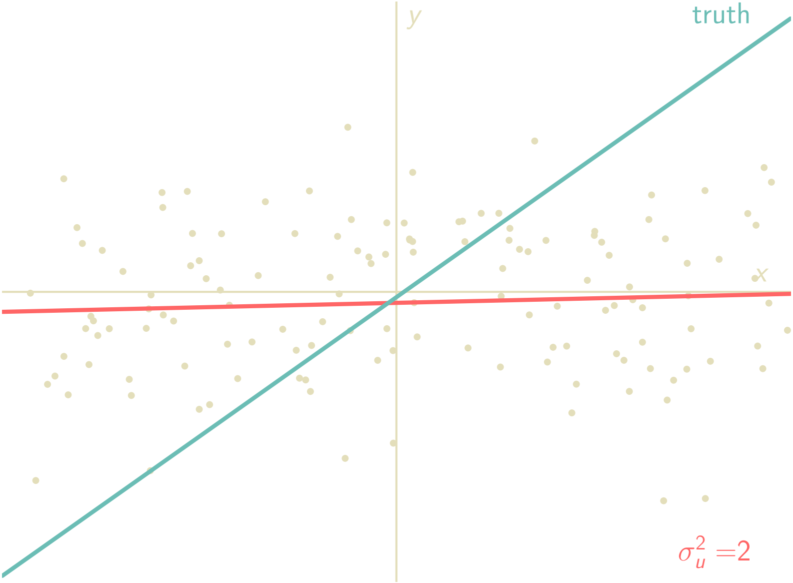

y
truth
x
$\sigma_u^2 = 0.3$


y
truth
x
$\sigma_u^2 = 0.4$


y
truth
x
$\sigma_u^2 = 0.6$

y
truth
x
$\sigma_u^2 = 0.8$


y
truth
x
$\sigma_u^2 = 1$


y
truth
x
$\sigma_u^2 = 2$

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.

observed missing

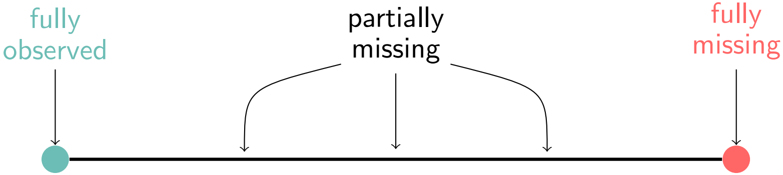
(fully) observed (fully) missing

(fully) observed (fully) missing

The false dichotomy of data.

fully
observed

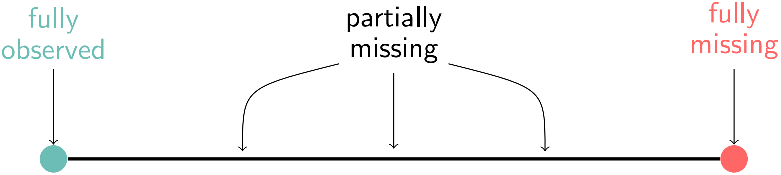
fully observed
fully missing

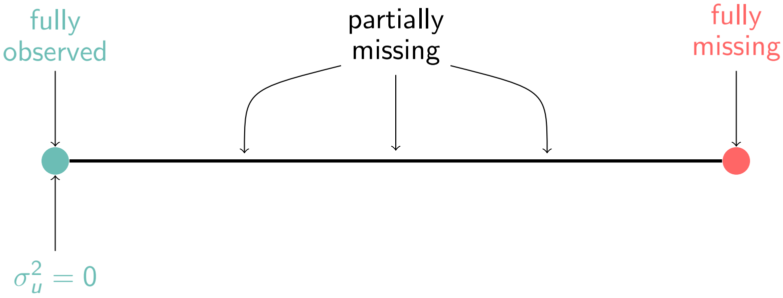

fully observed
partially missing
fully missing

But what is this continuum?

$$\overset{\substack{\text{observed}\\ \downarrow}}{x_i} = \overset{\substack{\text{latent}\\ \downarrow}}{x_i^*} + \overset{\substack{\text{measurement}\\ \text{error}\\ \downarrow}}{u_i}$$

$$u_i | x_i^* \sim \mathcal{N}(0, \underset{\substack{\uparrow\\ \text{measurement}\\ \text{error}\\ \text{variance}}}{\sigma_u^2})$$


fully observed
partially missing
fully missing


fully observed
partially missing
fully missing
$\sigma_u^2 = 0$

$$\overset{\text{observed}}{\downarrow}\quad\overset{\text{latent}}{\downarrow}\quad\overset{\text{measurement error}}{\downarrow}$$

$$x_i = x_i^* + u_i$$

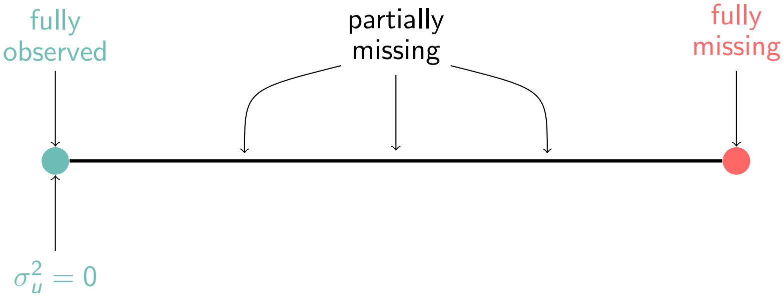
$$u_i \sim \mathcal{N}(0,0)$$

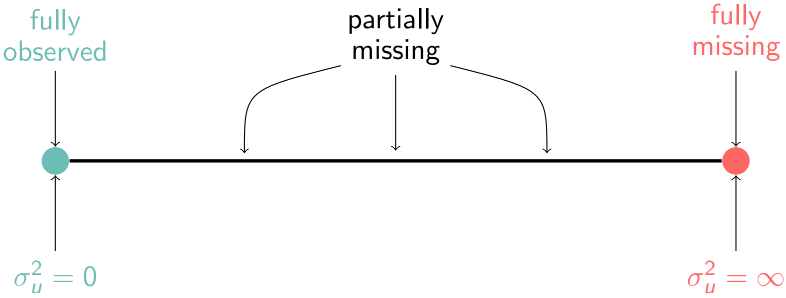
$$\uparrow$$

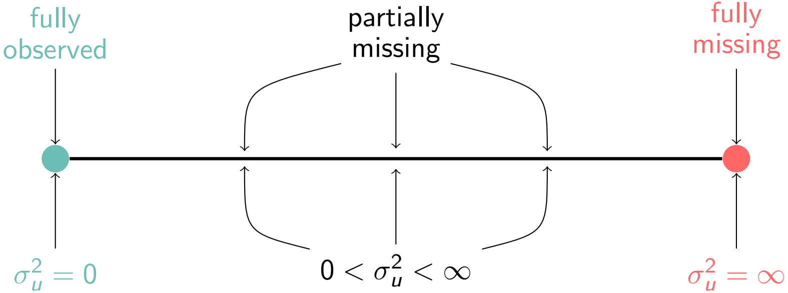
measurement
error
variance

$$\overset{\text{observed}}{x_i} \;=\; \overset{\text{latent}}{x_i^*} \;+\; \overset{\text{measurement error}}{0}$$

$$u_i \;\sim\; \mathcal{N}(0,\underset{\text{measurement error variance}}{0})$$


fully
observed
partially
missing
fully
missing
$\sigma_u^2 = 0$


fully observed
partially missing
fully missing
$\sigma_u^2 = 0$
$\sigma_u^2 = \infty$



fully observed

partially missing

fully missing

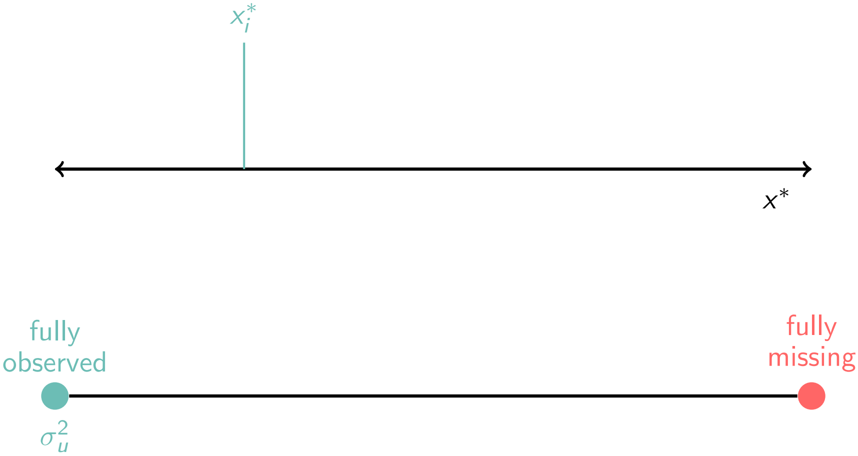
$\sigma_u^2 = 0$

$0 < \sigma_u^2 < \infty$

$\sigma_u^2 = \infty$

Missing data is the most
extreme case of measurement error.

x_i^*
x^*



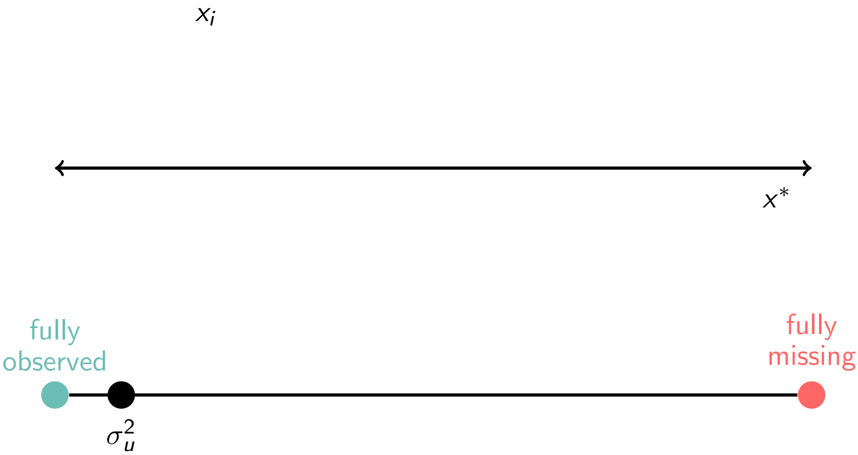
x_i^*

x^*

fully observed

fully missing

σ_u^2



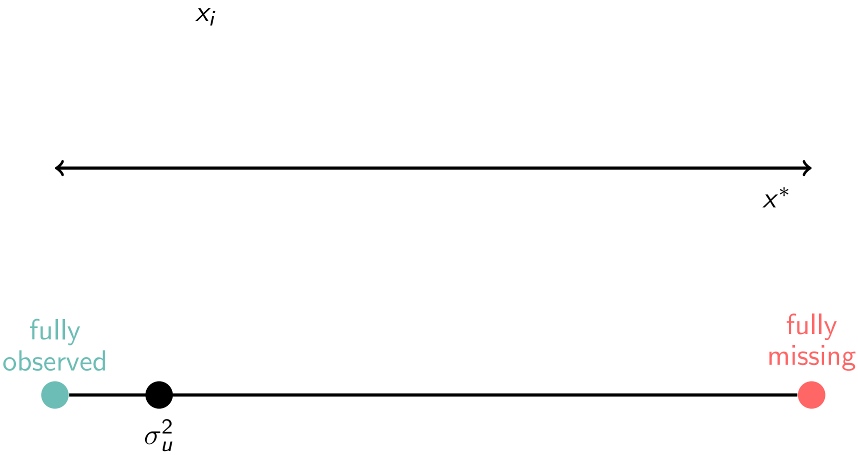
x_i

x^*

fully observed

fully missing

σ_u^2



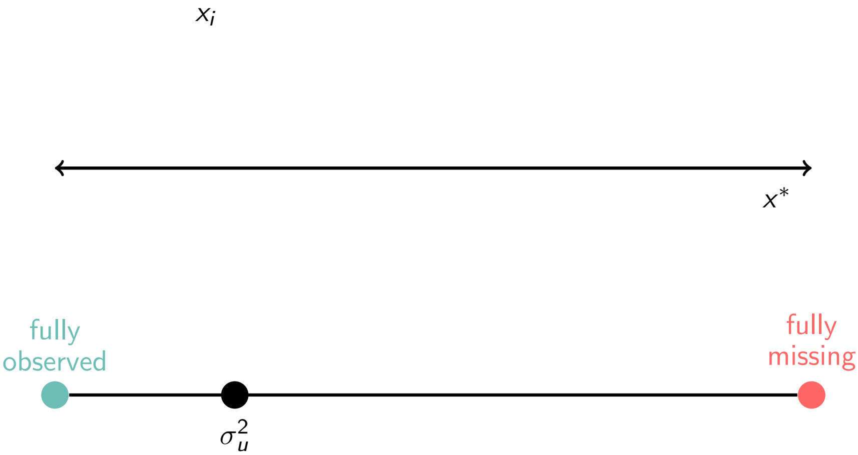
x_i

x^*

fully observed

fully missing

σ_u^2



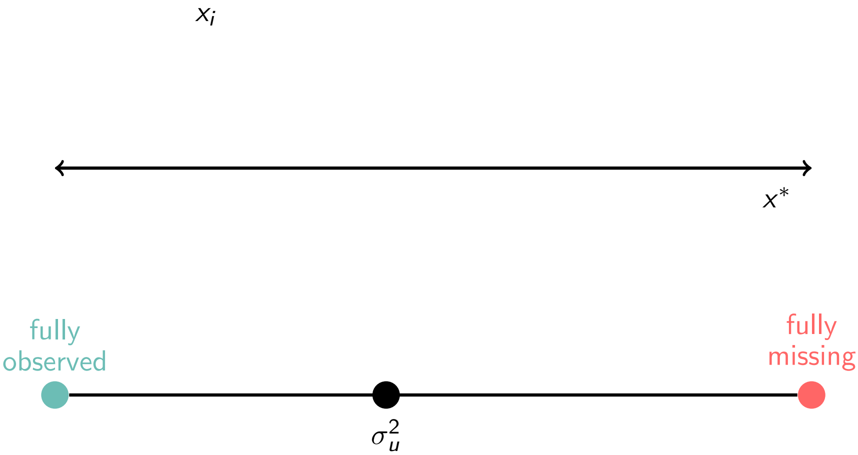
x_i

x^*

fully observed

fully missing

σ_u^2



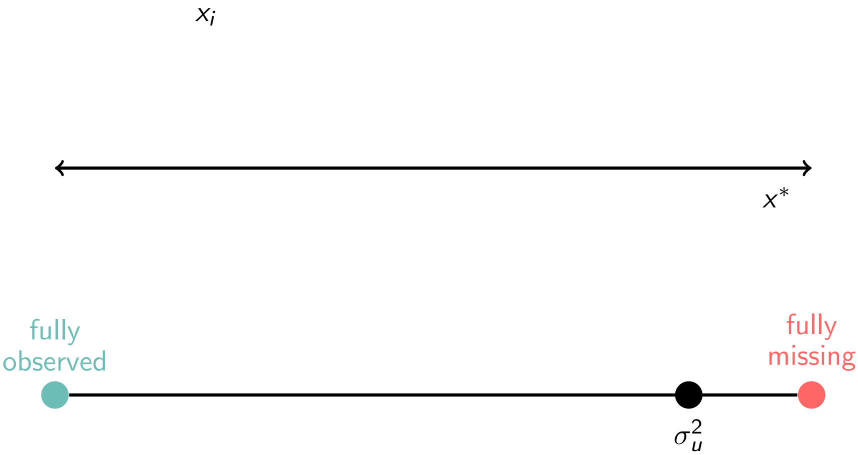
x_i

x^*

fully observed

fully missing

σ_u^2



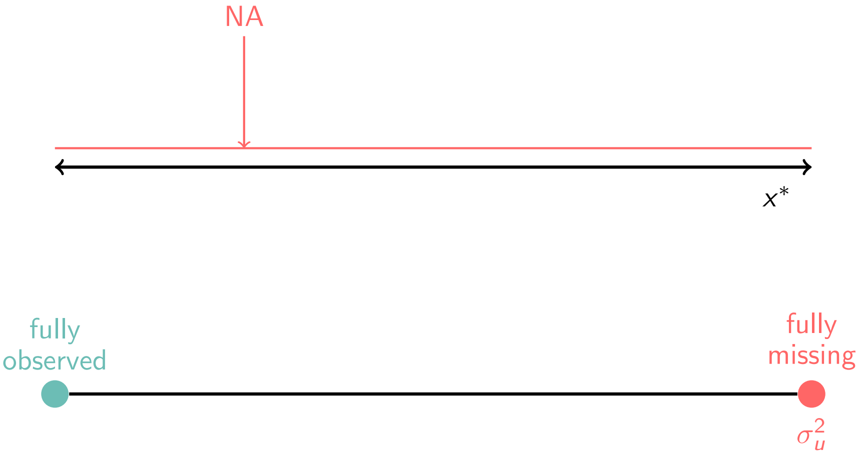
x_i

x^*

fully observed

fully missing

σ_u^2



NA

x^*

fully observed

fully missing

σ_u^2

Multiple imputation:

observed missing

Multiple imputation:

(fully) observed (fully) missing

Multiple overimputation:

x_i^* fully observed | | fully missing

Multiple overimputation:

x_i^* fully observed | partially missing | fully missing

Multiple overimputation:

x_i^*	fully observed perfectly measured	partially missing measured with error	fully missing infinite error

Multiple overimputation:

x_i^*	fully observed perfectly measured	partially missing measured with error	fully missing infinite error
$p(x_i \mid x_i^*)$	$\mathcal{N}(x_i^*, 0)$	$\mathcal{N}(x_i^*, \sigma_u^2)$	$\mathcal{N}(x_i^*, \infty)$

Multiple Overimputation
extends the multiple imputation framework
to correct for measurement error.

Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

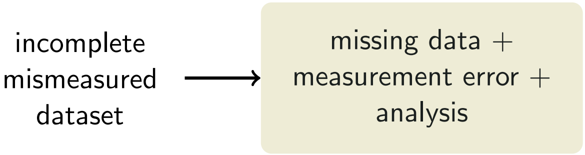
Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

incomplete
mismeasured
dataset

Missing Data and Measurement Error

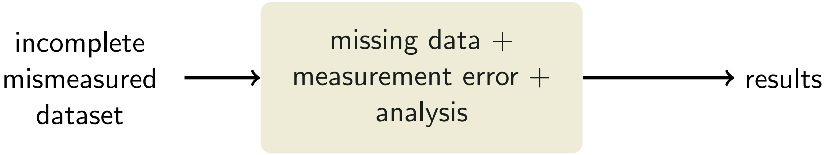
APPLICATION-SPECIFIC METHODS:



incomplete mismeasured dataset ⟶ missing data + measurement error + analysis

Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:



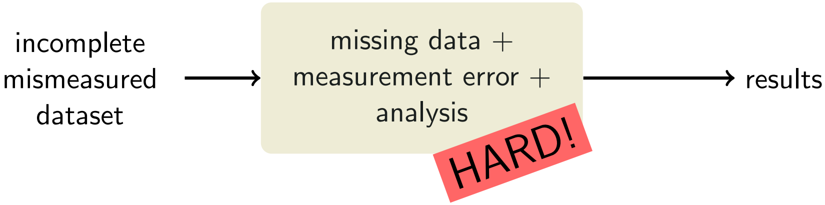
incomplete mismeasured dataset

missing data + measurement error + analysis

results

Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:



incomplete mismeasured dataset

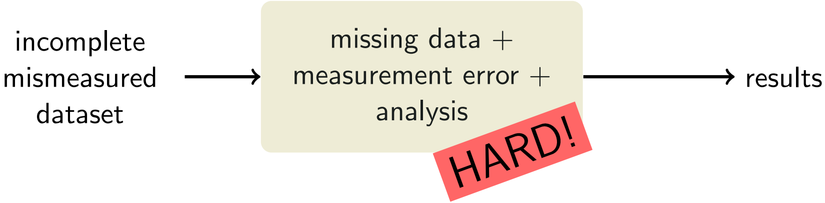
missing data + measurement error + analysis

HARD!

results

Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:



incomplete mismeasured dataset

missing data + measurement error + analysis

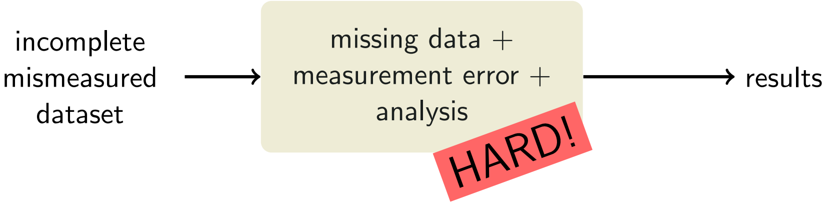
HARD!

results

MULTIPLE OVERIMPUTATION:

Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:



incomplete mismeasured dataset → missing data + measurement error + analysis → results

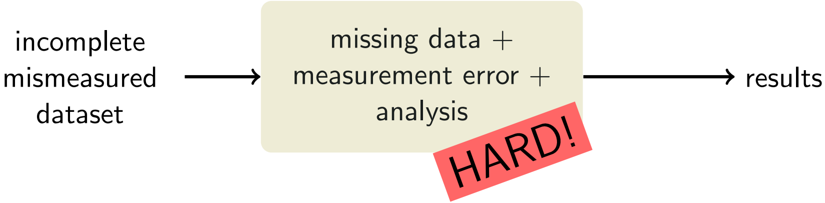
HARD!

MULTIPLE OVERIMPUTATION:

incomplete
mismeasured
dataset

Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:



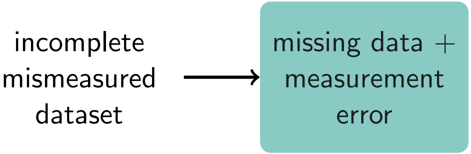
incomplete mismeasured dataset

missing data + measurement error + analysis

HARD!

results

MULTIPLE OVERIMPUTATION:

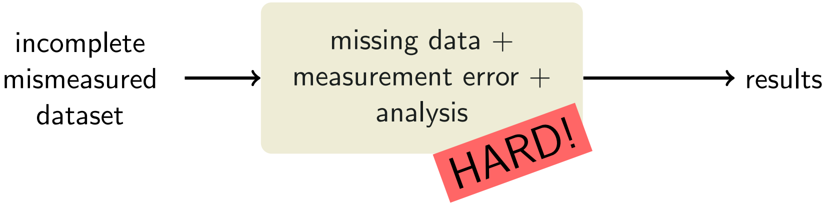


incomplete mismeasured dataset

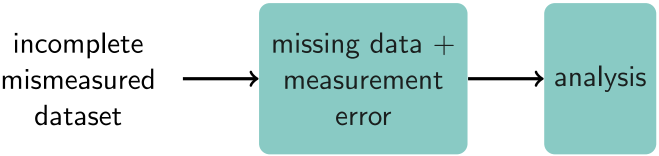
missing data + measurement error

Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

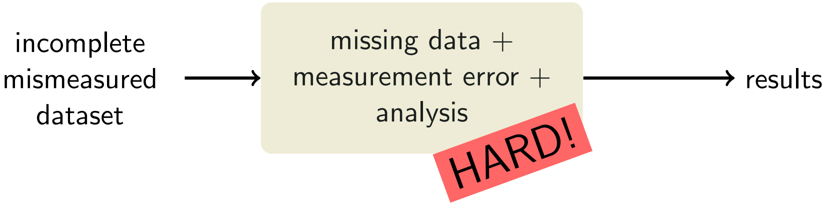

incomplete mismeasured dataset
missing data + measurement error + analysis
HARD!
results

MULTIPLE OVERIMPUTATION:

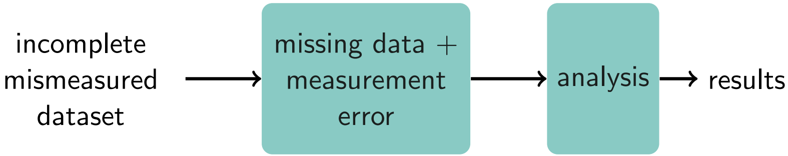

incomplete mismeasured dataset
missing data + measurement error
analysis

Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:

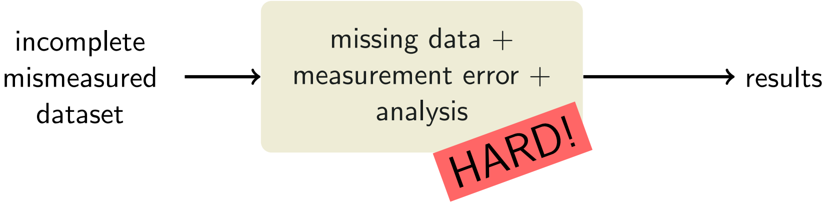


MULTIPLE OVERIMPUTATION:



Missing Data and Measurement Error

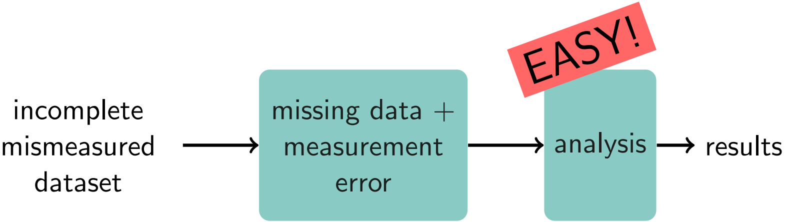
APPLICATION-SPECIFIC METHODS:



incomplete mismeasured dataset → missing data + measurement error + analysis → results

HARD!

MULTIPLE OVERIMPUTATION:

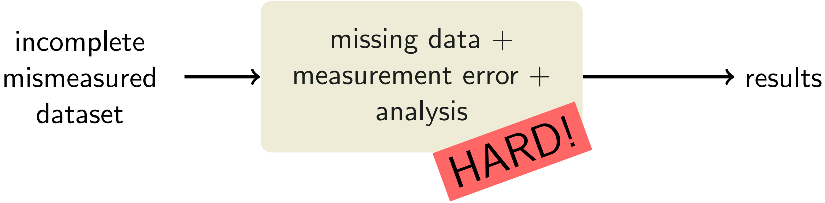


incomplete mismeasured dataset → missing data + measurement error → analysis → results

EASY!

Missing Data and Measurement Error

APPLICATION-SPECIFIC METHODS:



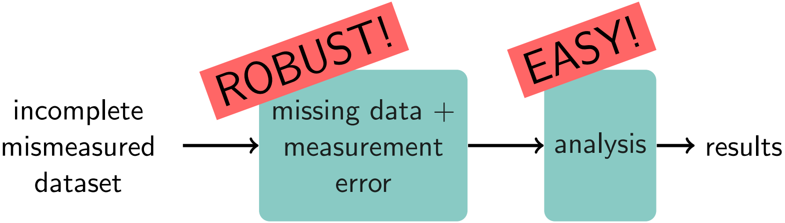
incomplete mismeasured dataset

missing data + measurement error + analysis

HARD!

results

MULTIPLE OVERIMPUTATION:



incomplete mismeasured dataset

ROBUST!

missing data + measurement error

EASY!

analysis

results

What MO allows you to do:

What MO allows you to do:

social science.

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	$\approx$9	6	6.23	5.92
2	LIBERIA	NA	3	NA	NA
3	SIERRA LEONE	$\approx$3	3	6.60	NA
4	GHANA	$\approx$9	6	6.86	12.68
5	TOGO	NA	5	6.27	17.34
6	CAMEROON	$\approx$6	5	6.93	15.47
7	NIGERIA	$\approx$5	7	6.88	17.46
8	GABON	$\approx$6	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	≈9	6	6.23	5.92
2	LIBERIA	⌒	3	⌒	⌒
3	SIERRA LEONE	≈3	3	6.60	⌒
4	GHANA	≈9	6	6.86	12.68
5	TOGO	⌒	5	6.27	17.34
6	CAMEROON	≈6	5	6.93	15.47
7	NIGERIA	≈5	7	6.88	17.46
8	GABON	≈6	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	∧	6	6.23	5.92
2	LIBERIA	⌒	3	⌒	⌒
3	SIERRA LEONE	∧	3	6.60	⌒
4	GHANA	∧	6	6.86	12.68
5	TOGO	⌒	5	6.27	17.34
6	CAMEROON	∧	5	6.93	15.47
7	NIGERIA	∧	7	6.88	17.46
8	GABON	∧	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	⋏	6	6.23	5.92
2	LIBERIA	⌒	3	⌒	⌒
3	SIERRA LEONE	⋏	3	6.60	⌒
4	GHANA	⋏	6	6.86	12.68
5	TOGO	⌒	5	6.27	17.34
6	CAMEROON	⋏	5	6.93	15.47
7	NIGERIA	⋏	7	6.88	17.46
8	GABON	⋏	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	⋏	6	6.23	5.92
2	LIBERIA	⌒	3	⌒	⌒
3	SIERRA LEONE	⋏	3	6.60	⌒
4	GHANA	⋏	6	6.86	12.68
5	TOGO	⌒	5	6.27	17.34
6	CAMEROON	⋏	5	6.93	15.47
7	NIGERIA	⋏	7	6.88	17.46
8	GABON	⋏	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	⋏	6	6.23	5.92
2	LIBERIA	⌒	3	⌒	⌒
3	SIERRA LEONE	⋏	3	6.60	⌒
4	GHANA	⋏	6	6.86	12.68
5	TOGO	⌒	5	6.27	17.34
6	CAMEROON	⋏	5	6.93	15.47
7	NIGERIA	⋏	7	6.88	17.46
8	GABON	⋏	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	⋏	6	6.23	5.92
2	LIBERIA	⌒	3	⌒	⌒
3	SIERRA LEONE	⋏	3	6.60	⌒
4	GHANA	⋏	6	6.86	12.68
5	TOGO	⌒	5	6.27	17.34
6	CAMEROON	⋏	5	6.93	15.47
7	NIGERIA	⋏	7	6.88	17.46
8	GABON	⋏	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	⋏	6	6.23	5.92
2	LIBERIA	⌒	3	⌒	⌒
3	SIERRA LEONE	⋏	3	6.60	⌒
4	GHANA	⋏	6	6.86	12.68
5	TOGO	⌒	5	6.27	17.34
6	CAMEROON	⋏	5	6.93	15.47
7	NIGERIA	⋏	7	6.88	17.46
8	GABON	⋏	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	∧	6	6.23	5.92
2	LIBERIA	⌒	3	⌒	⌒
3	SIERRA LEONE	∧	3	6.60	⌒
4	GHANA	∧	6	6.86	12.68
5	TOGO	⌒	5	6.27	17.34
6	CAMEROON	∧	5	6.93	15.47
7	NIGERIA	∧	7	6.88	17.46
8	GABON	∧	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	∧	6	6.23	5.92
2	LIBERIA	⌒	3	⌒	⌒
3	SIERRA LEONE	∧	3	6.60	⌒
4	GHANA	∧	6	6.86	12.68
5	TOGO	⌒	5	6.27	17.34
6	CAMEROON	∧	5	6.93	15.47
7	NIGERIA	∧	7	6.88	17.46
8	GABON	∧	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	∧	6	6.23	5.92
2	LIBERIA	⌒	3	⌒	⌒
3	SIERRA LEONE	∧	3	6.60	⌒
4	GHANA	∧	6	6.86	12.68
5	TOGO	⌒	5	6.27	17.34
6	CAMEROON	∧	5	6.93	15.47
7	NIGERIA	∧	7	6.88	17.46
8	GABON	∧	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	∧	6	6.23	5.92
2	LIBERIA	⌒	3	⌒	⌒
3	SIERRA LEONE	∧	3	6.60	⌒
4	GHANA	∧	6	6.86	12.68
5	TOGO	⌒	5	6.27	17.34
6	CAMEROON	∧	5	6.93	15.47
7	NIGERIA	∧	7	6.88	17.46
8	GABON	∧	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	∧	6	6.23	5.92
2	LIBERIA	⌒	3	⌒	⌒
3	SIERRA LEONE	∧	3	6.60	⌒
4	GHANA	∧	6	6.86	12.68
5	TOGO	⌒	5	6.27	17.34
6	CAMEROON	∧	5	6.93	15.47
7	NIGERIA	∧	7	6.88	17.46
8	GABON	∧	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	∧	6	6.23	5.92
2	LIBERIA	⌒	3	⌒	⌒
3	SIERRA LEONE	∧	3	6.60	⌒
4	GHANA	∧	6	6.86	12.68
5	TOGO	⌒	5	6.27	17.34
6	CAMEROON	∧	5	6.93	15.47
7	NIGERIA	∧	7	6.88	17.46
8	GABON	∧	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	∧	6	6.23	5.92
2	LIBERIA	⌒	3	⌒	⌒
3	SIERRA LEONE	∧	3	6.60	⌒
4	GHANA	∧	6	6.86	12.68
5	TOGO	⌒	5	6.27	17.34
6	CAMEROON	∧	5	6.93	15.47
7	NIGERIA	∧	7	6.88	17.46
8	GABON	∧	8	8.19	16.97

	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	∧	6	6.23	5.92
2	LIBERIA	⌒	3	⌒	⌒
3	SIERRA LEONE	∧	3	6.60	⌒
4	GHANA	∧	6	6.86	12.68
5	TOGO	⌒	5	6.27	17.34
6	CAMEROON	∧	5	6.93	15.47
7	NIGERIA	∧	7	6.88	17.46
8	GABON	∧	8	8.19	16.97

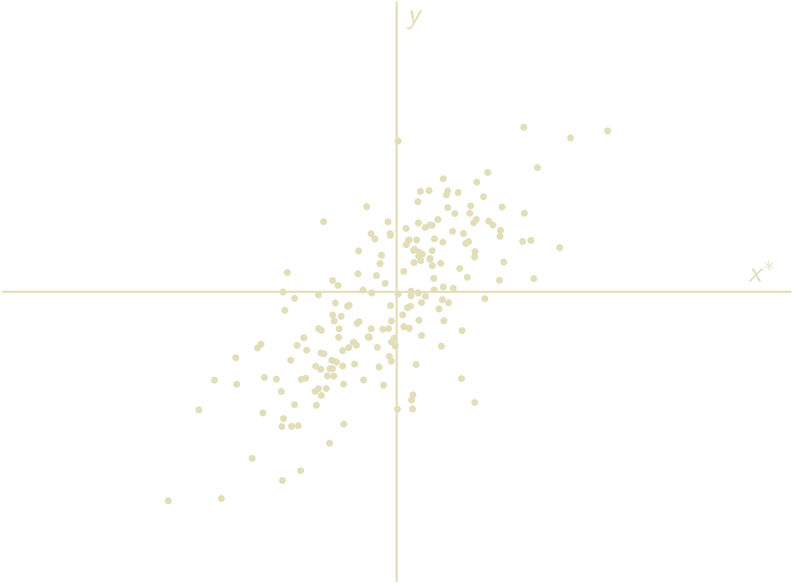
	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	∧	6	6.23	5.92
2	LIBERIA	⌒	3	⌒	⌒
3	SIERRA LEONE	∧	3	6.60	⌒
4	GHANA	∧	6	6.86	12.68
5	TOGO	⌒	5	6.27	17.34
6	CAMEROON	∧	5	6.93	15.47
7	NIGERIA	∧	7	6.88	17.46
8	GABON	∧	8	8.19	16.97

Run whatever analysis model you wanted to run.

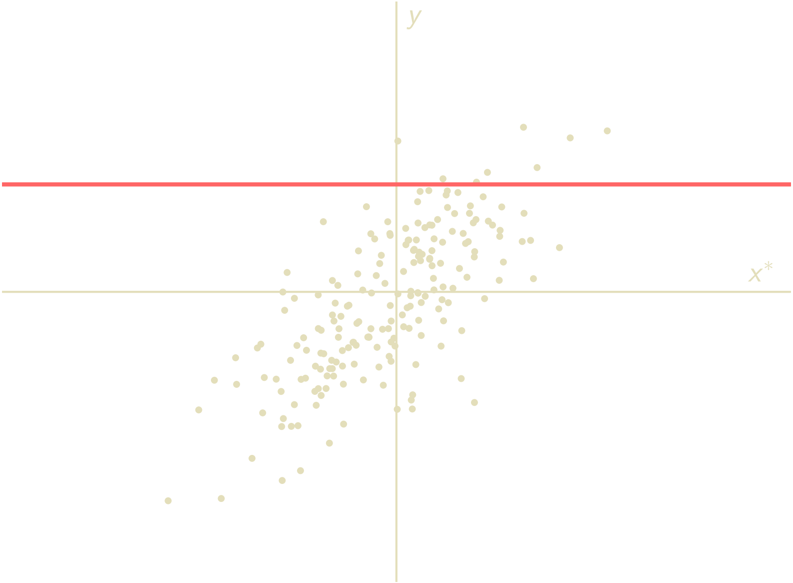
Run whatever analysis model you wanted to run.
(×5)

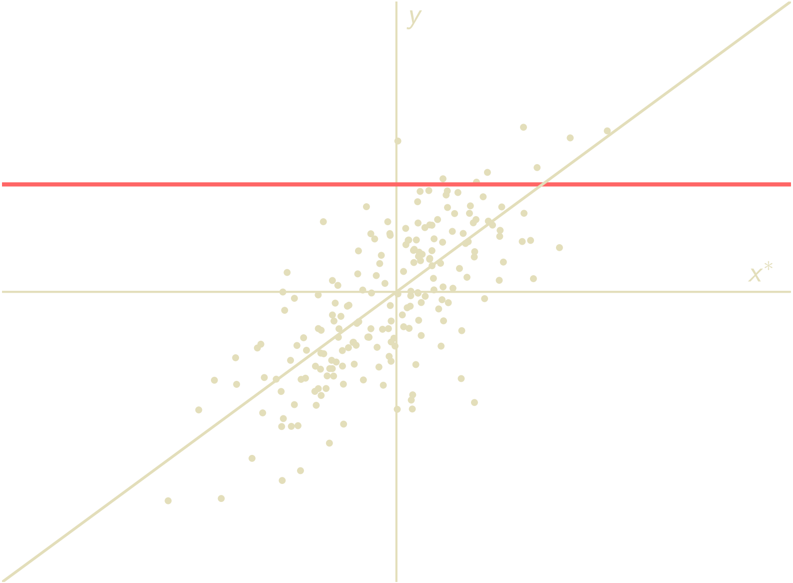
But how does it work?

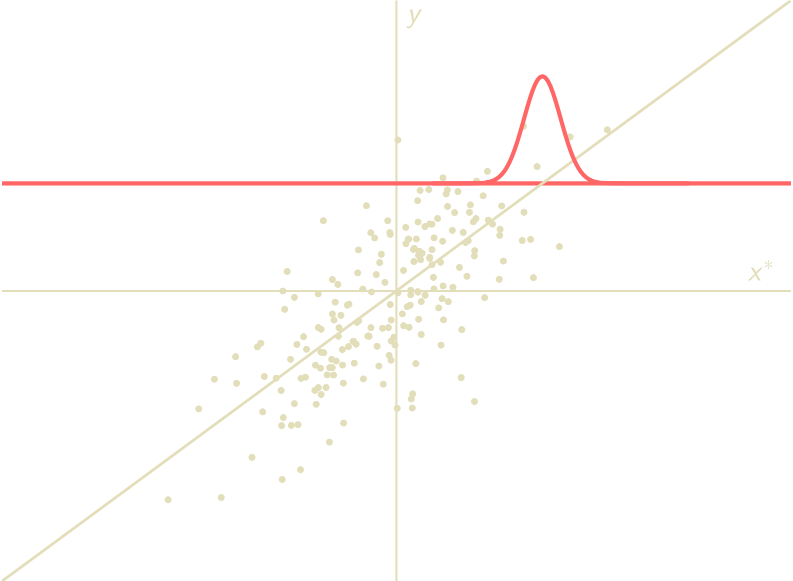
Let's look at the extreme case first.

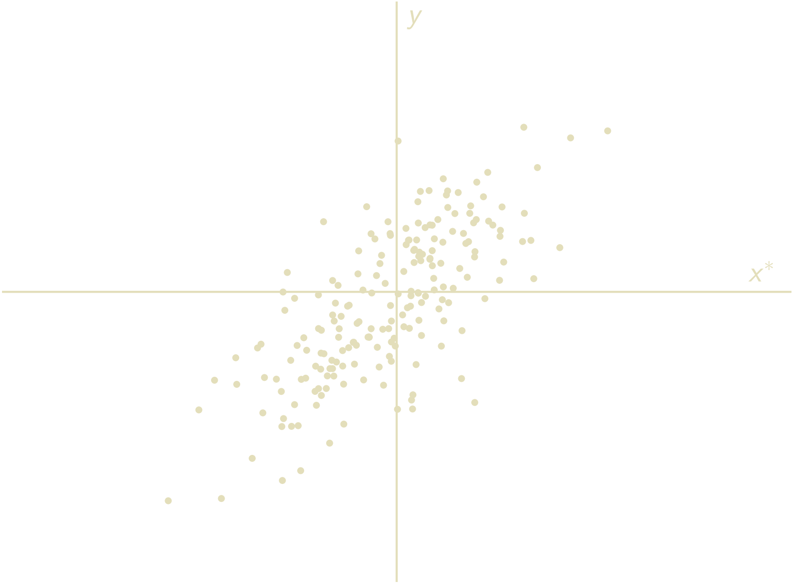


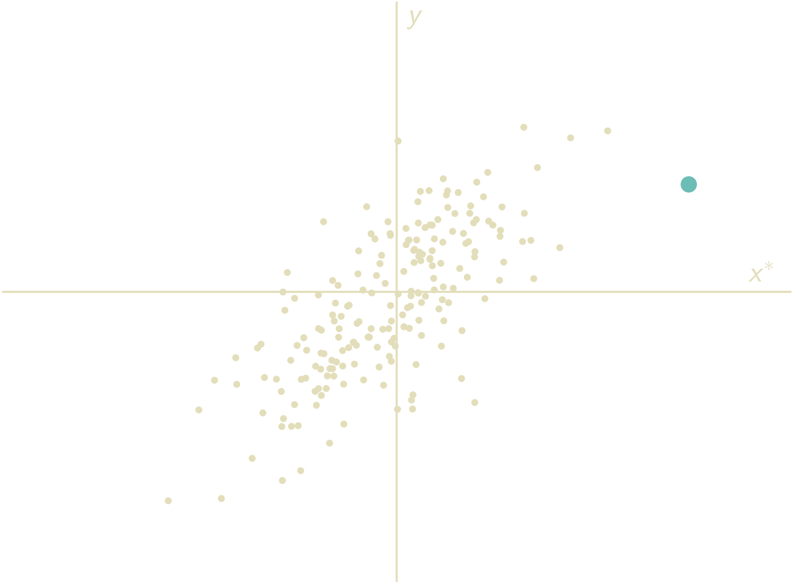
y
x^*

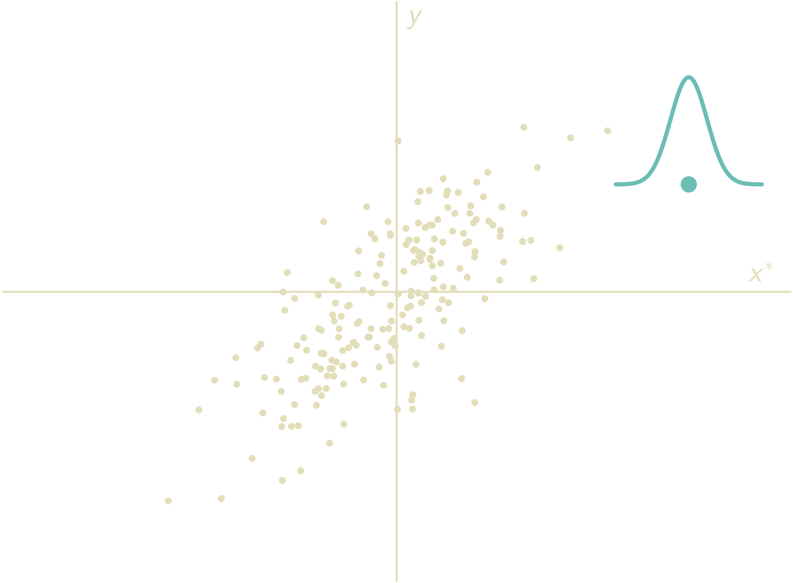

y
x^*

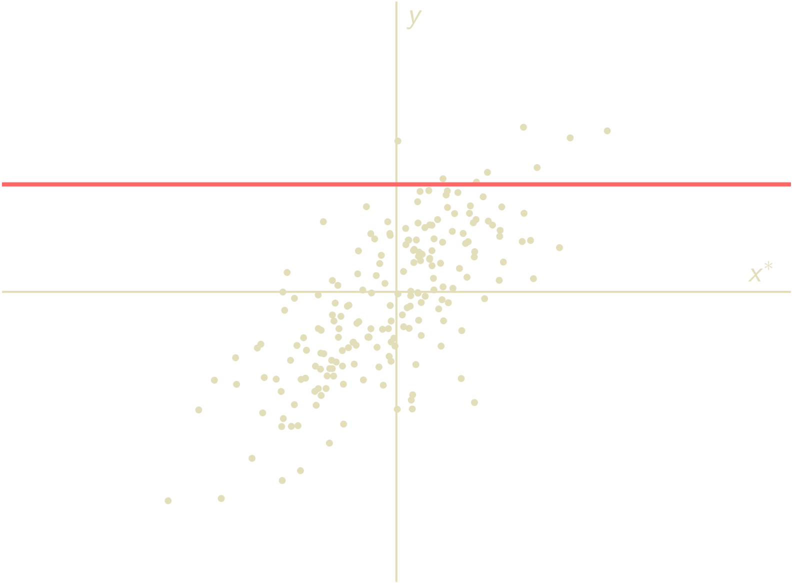

y
x^*

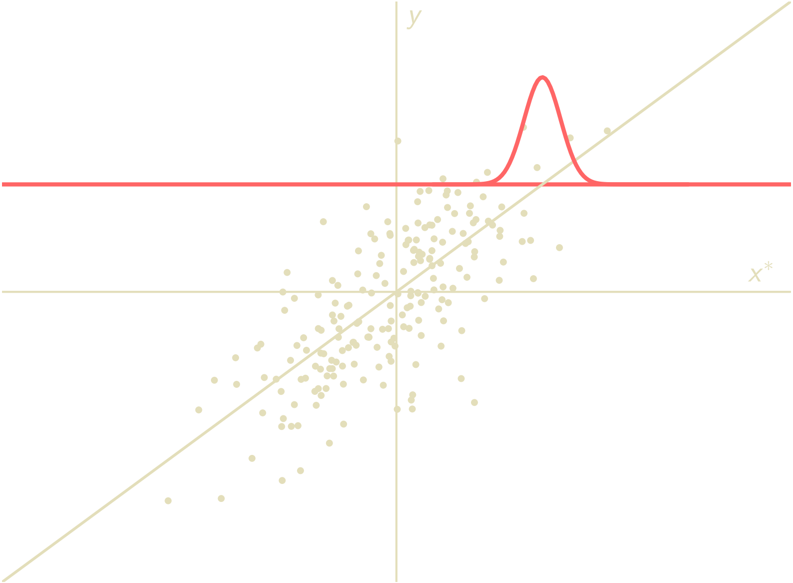

y
x^*

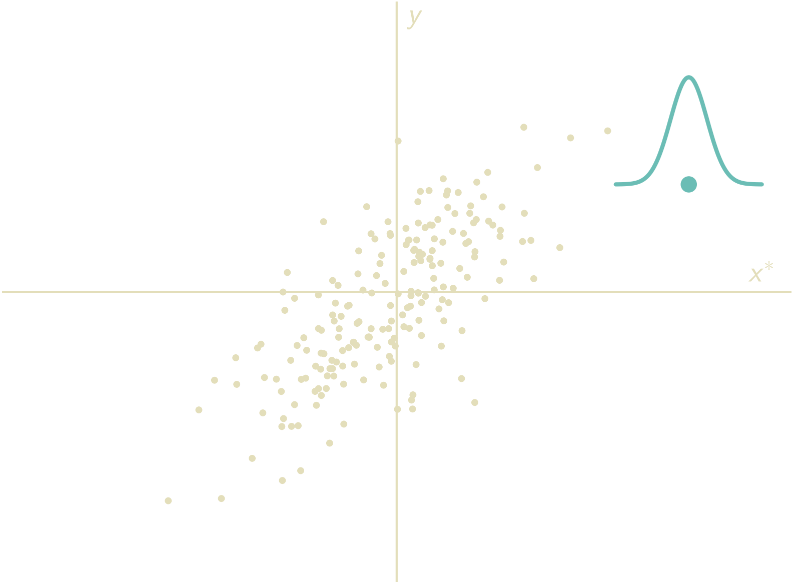

y
x^*

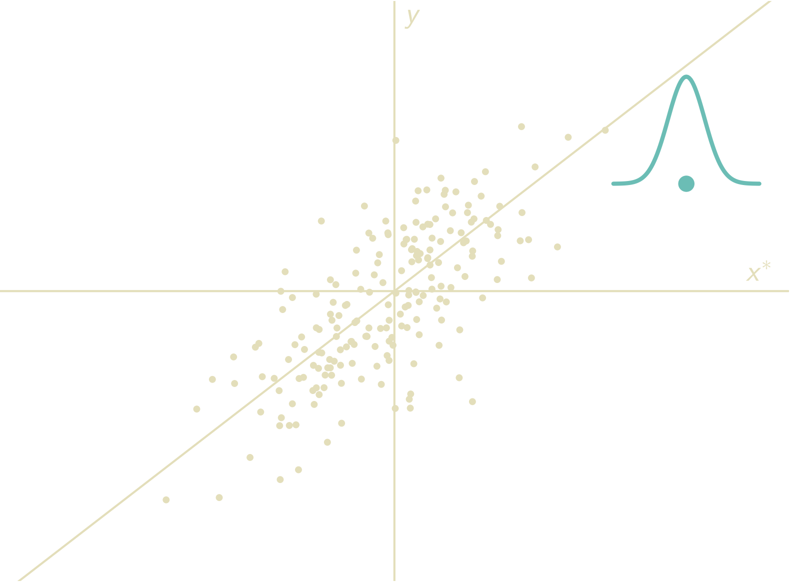

y
x^*

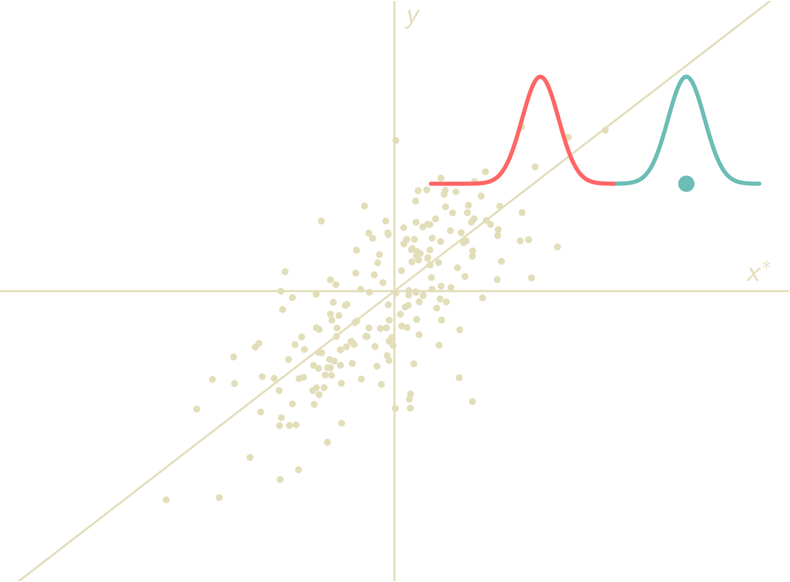

y
x^*

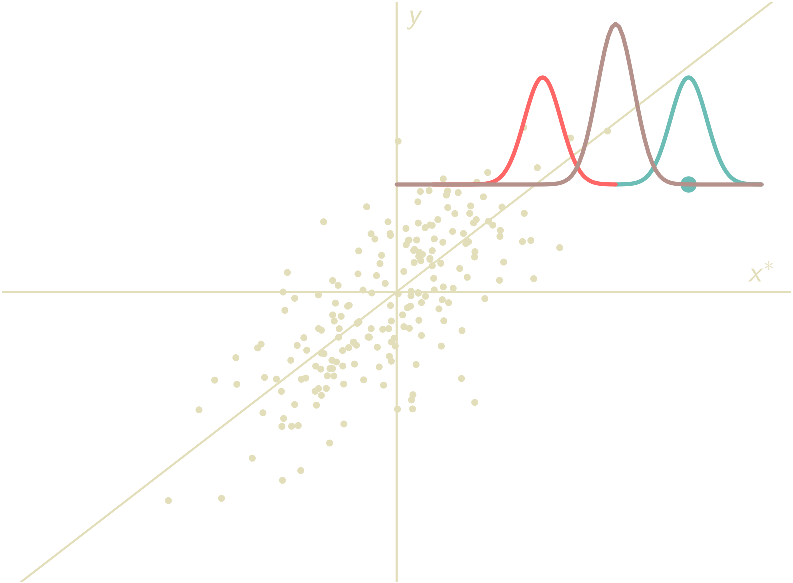

y
x^*


y
x^*


y
x^*


y
x^*


y
x*


y
x*

ARBITRARY PATTERNS OF MISMEASUREMENT & MISSINGNESS:

ARBITRARY PATTERNS OF MISMEASUREMENT & MISSINGNESS:

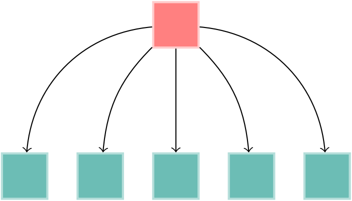
	country	polityiv	f-house	log-gdppc	primary
1	BUKINA FASO	4	≈6	6.23	5.92
2	LIBERIA	NA	3	NA	NA
3	SIERRA LEONE	3	3	6.60	NA
4	GHANA	≈9	6	6.86	12.68
5	TOGO	NA	5	6.27	17.34
6	CAMEROON	≈6	5	6.93	15.47
7	NIGERIA	≈5	7	≈6.88	17.46
8	GABON	≈6	8	≈8.19	≈16.97

The Multiple Imputation Scheme

The Multiple Imputation Scheme

incomplete data

The Multiple Imputation Scheme

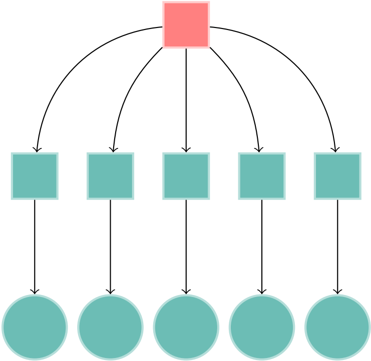


incomplete data

posterior predictive draws

imputed datasets

The Multiple Imputation Scheme



incomplete data

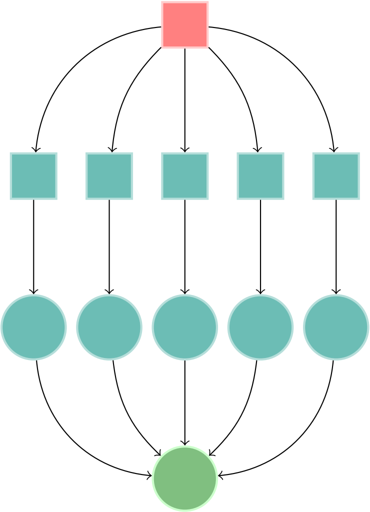
posterior predictive draws

imputed datasets

analysis

separate results

The Multiple Imputation Scheme



incomplete data

posterior predictive draws

imputed datasets

analysis

separate results

combination

final results

(1)
Mismeasured at random (MMAR).

(2)

You have to know how things were mismeasured.

(2)
You have to know $f(x_i|x_i^*)$.

(3^*)

Measurement error and ideal data are statistically dual.

OUR SPECIFIC MODEL

OUR SPECIFIC MODEL

ideal data

$$(y_i, x_i^*) \sim \mathcal{MVN}(\mu, \Sigma)$$

OUR SPECIFIC MODEL

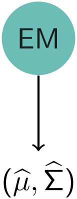
ideal data

$$(y_i, x_i^*) \sim \mathcal{MVN}(\mu, \Sigma)$$

$$x_i \sim \mathcal{N}(x_i^*, \sigma_u^2)$$

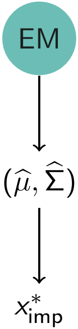
measurement error

EM



EM

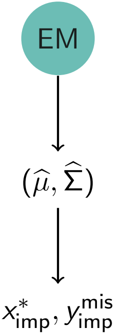
$(\widehat{\mu}, \widehat{\Sigma})$



EM

$(\widehat{\mu}, \widehat{\Sigma})$

x^*_{imp}



EM

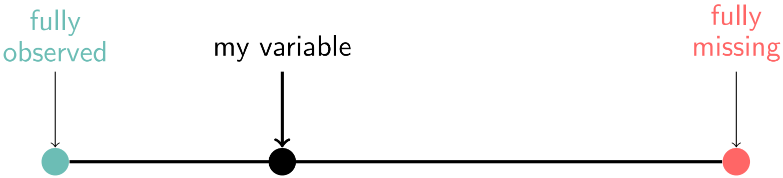
$(\widehat{\mu}, \widehat{\Sigma})$

$x^*_{\text{imp}}, y^{\text{mis}}_{\text{imp}}$

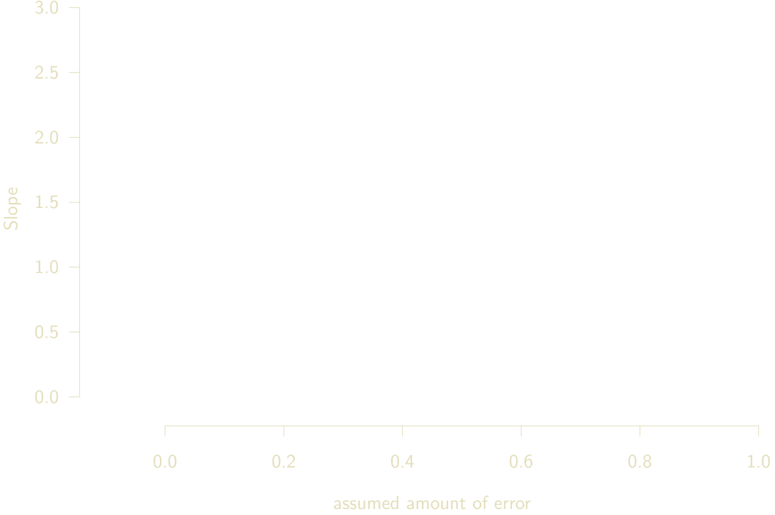
CHOOSE A VALUE OF σ_u^2

fully observed
fully missing

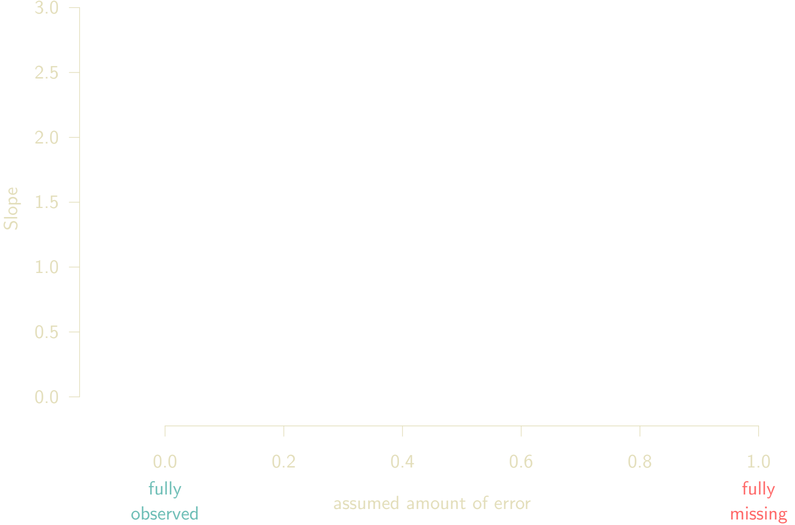
CHOOSE A VALUE OF σ_u^2

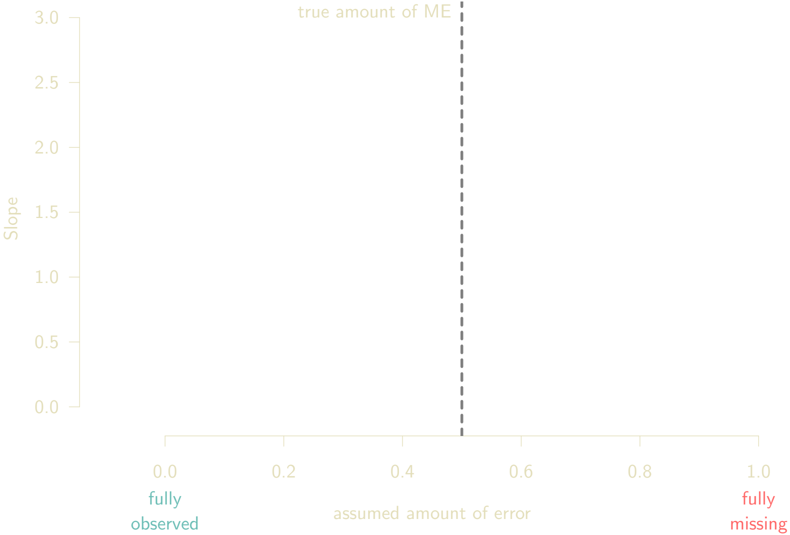


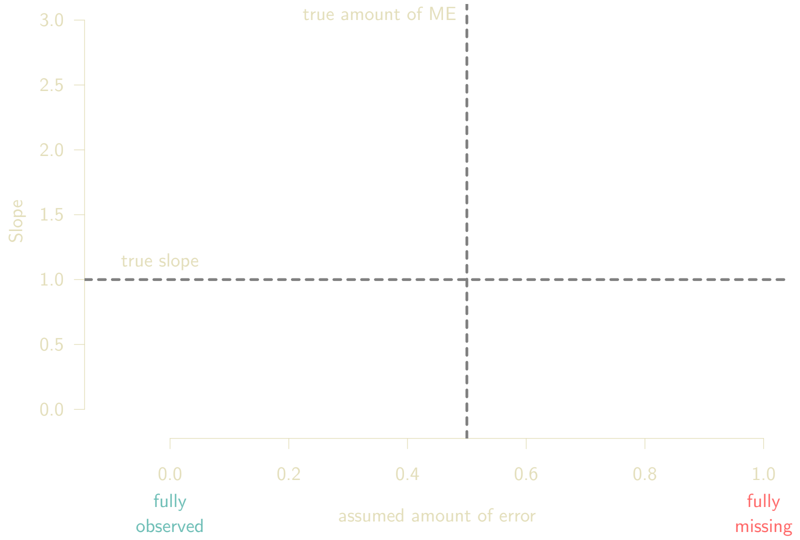
Some simulations.

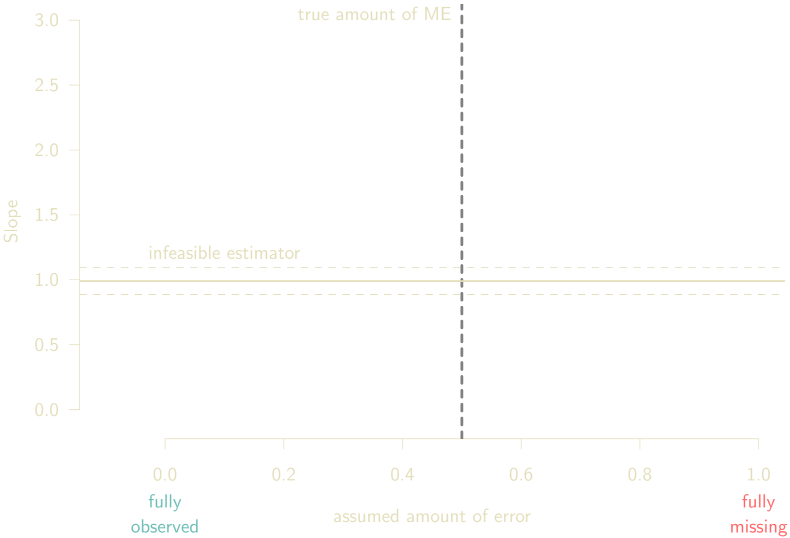

3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
0.0
0.2
0.4
0.6
0.8
1.0
assumed amount of error

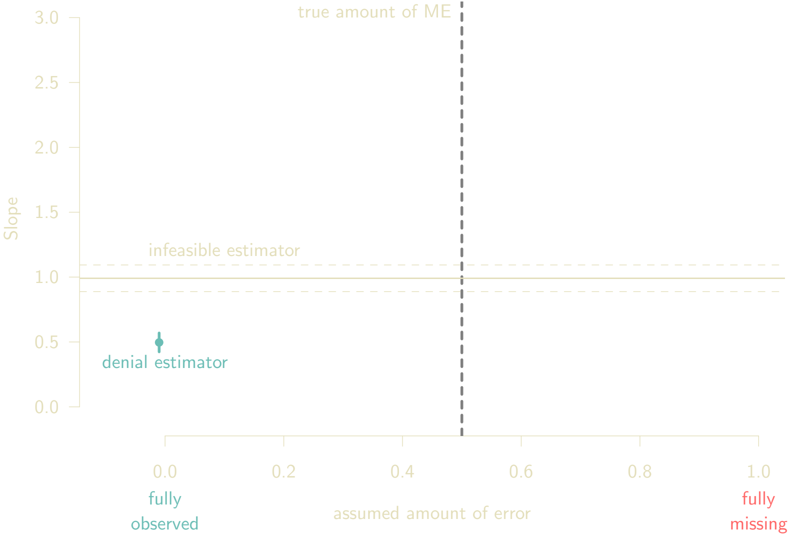
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
0.0
0.2
0.4
0.6
0.8
1.0
fully
observed
assumed amount of error

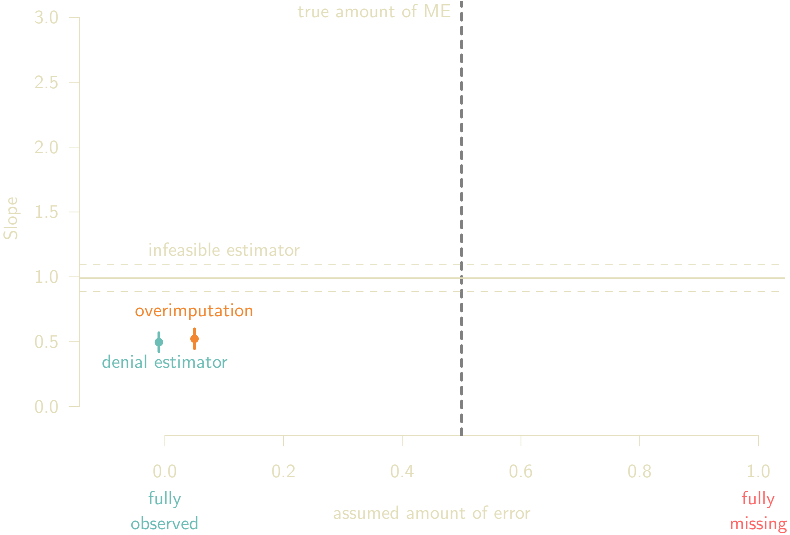

3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
0.0
0.2
0.4
0.6
0.8
1.0
fully
observed
assumed amount of error
fully
missing

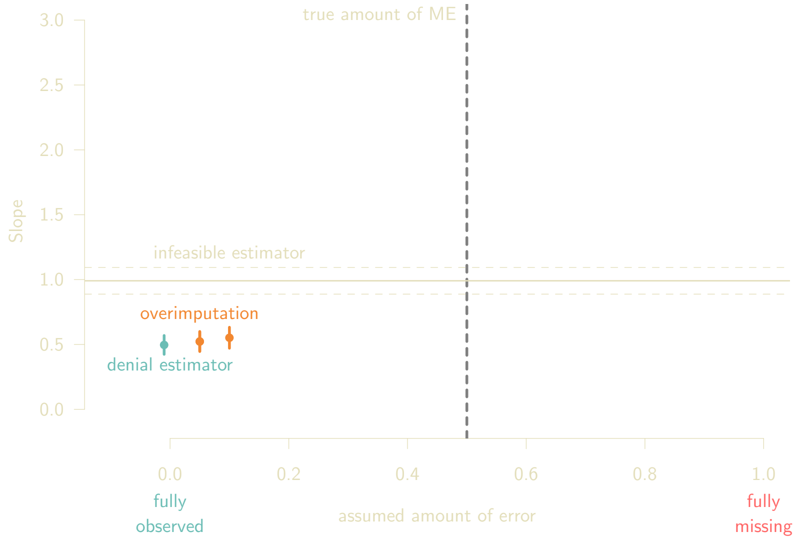

true amount of ME
Slope
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
fully observed
assumed amount of error
fully missing

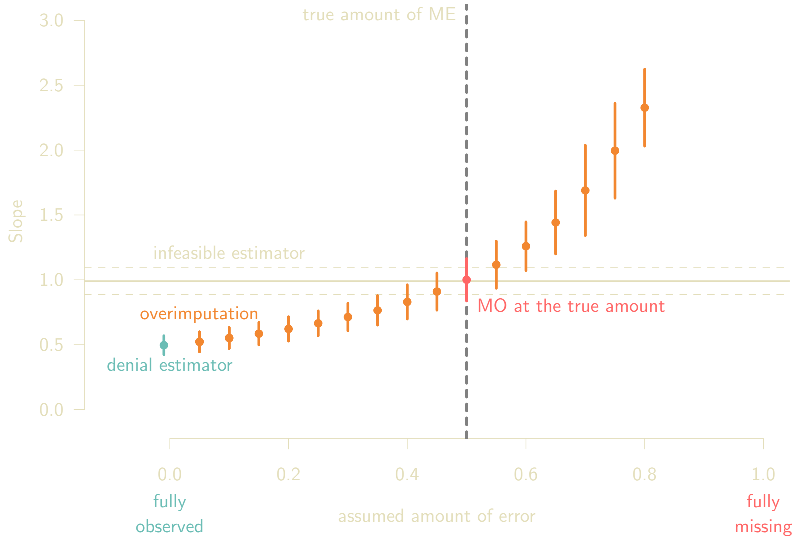

true amount of ME
3.0
2.5
2.0
1.5
Slope
true slope
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
fully
observed
assumed amount of error
fully
missing

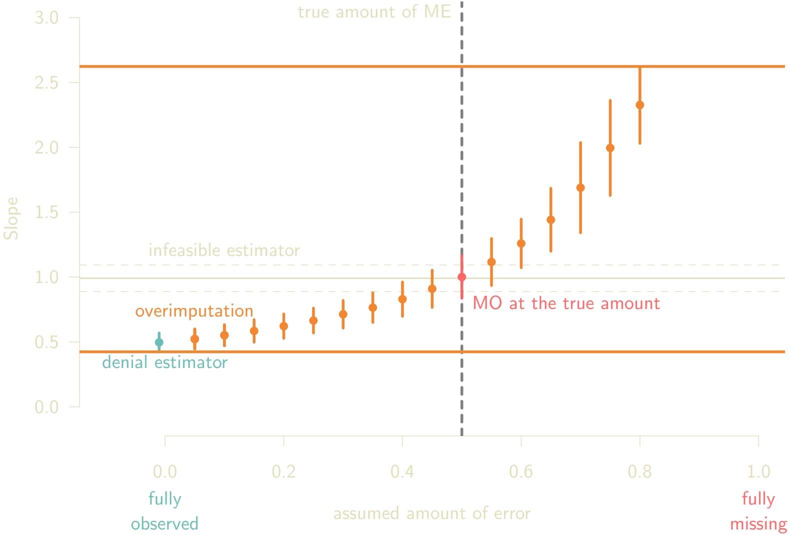

true amount of ME
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
infeasible estimator
0.0
0.2
0.4
0.6
0.8
1.0
fully
observed
assumed amount of error
fully
missing


true amount of ME
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
infeasible estimator
denial estimator
0.0
0.2
0.4
0.6
0.8
1.0
fully
observed
assumed amount of error
fully
missing

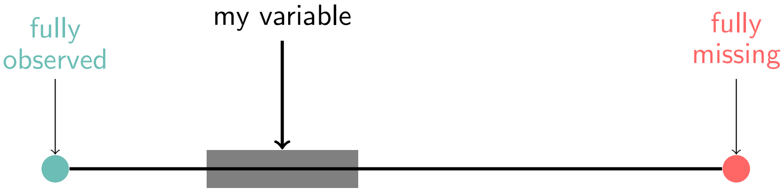

true amount of ME
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
infeasible estimator
overimputation
denial estimator
0.0
0.2
0.4
0.6
0.8
1.0
fully
observed
assumed amount of error
fully
missing

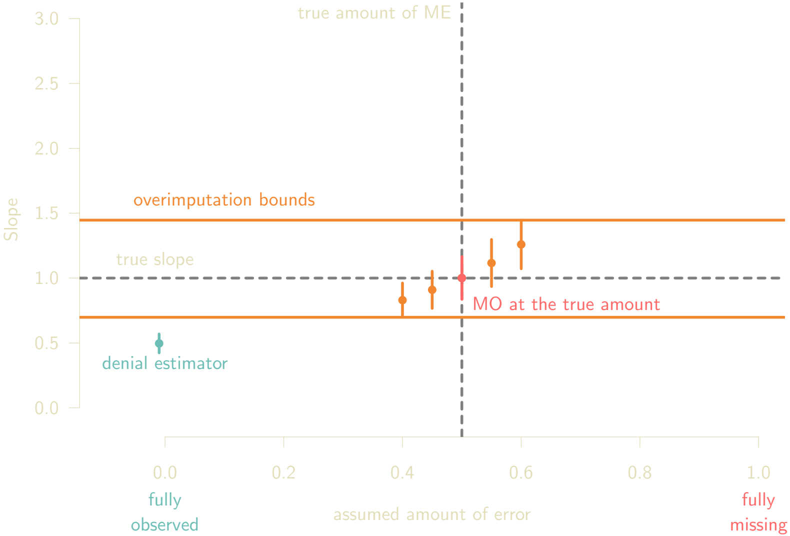

true amount of ME
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
infeasible estimator
overimputation
denial estimator
0.0
0.2
0.4
0.6
0.8
1.0
fully
observed
assumed amount of error
fully
missing


true amount of ME
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
infeasible estimator
overimputation
MO at the true amount
denial estimator
0.0
0.2
0.4
0.6
0.8
1.0
fully
observed
assumed amount of error
fully
missing


true amount of ME
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
infeasible estimator
overimputation
MO at the true amount
denial estimator
0.0
0.2
0.4
0.6
0.8
1.0
fully
observed
assumed amount of error
fully
missing

CHOOSE A RANGE OF σ_u^2




true amount of ME
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Slope
overimputation bounds
true slope
MO at the true amount
denial estimator
0.0
0.2
0.4
0.6
0.8
1.0
fully
observed
assumed amount of error
fully
missing

	Missing Data	Measurement Error
20 Years Ago		
Today		

	Missing Data	Measurement Error
20 Years Ago	TAILORED METHODS:	TAILORED METHODS:
Today		

	Missing Data	Measurement Error
20 Years Ago	TAILORED METHODS: Model dependent	TAILORED METHODS: Model dependent
Today		

	Missing Data	Measurement Error
20 Years Ago	TAILORED METHODS: Model dependent Difficult to implement	TAILORED METHODS: Model dependent Difficult to implement
Today		

	Missing Data	Measurement Error
20 Years Ago	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions
Today		

	Missing Data	Measurement Error
20 Years Ago	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions
Today	MULTIPLE IMPUTATION:	

	Missing Data	Measurement Error
20 Years Ago	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions
Today	MULTIPLE IMPUTATION: Broadly applicable	

	Missing Data	Measurement Error
20 Years Ago	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions
Today	MULTIPLE IMPUTATION: Broadly applicable Easy to implement	

	Missing Data	Measurement Error
20 Years Ago	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions
Today	MULTIPLE IMPUTATION: Broadly applicable Easy to implement Widely used.	

	Missing Data	Measurement Error
20 Years Ago	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions	TAILORED METHODS: Model dependent Difficult to implement Dubious assumptions
Today	MULTIPLE IMPUTATION: Broadly applicable Easy to implement Widely used.	MULTIPLE OVERIMPUTATION

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?

◂ Return

How Bad Is Listwise Deletion?

◂ Return

Goal: estimate β_1, where X_2 has λ missing values (y, X_1 are fully observed).

How Bad Is Listwise Deletion?

◂ Return

Goal: estimate β_1, where X_2 has λ missing values (y, X_1 are fully observed).

$$E(y) = X_1\beta_1 + X_2\beta_2$$

How Bad Is Listwise Deletion?

◂ Return

Goal: estimate β_1, where X_2 has λ 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?

◂ Return

Goal: estimate β_1, where X_2 has λ 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.

How Bad Is Listwise Deletion?

◂ Return

Goal: estimate β_1, where X_2 has λ 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 β_1 by b_1^O, the slope from regressing y on X_1.

How Bad Is Listwise Deletion?

◂ Return

Goal: estimate β_1, where X_2 has λ 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 β_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 β_1 as b_1^L, the coefficient on X_1 when regressing y on X_1 and X_2.

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.

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 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.

$$\begin{aligned}\text{MSE}(\hat{a}) &= E[(\hat{a} - a)^2] \\ &= V(\hat{a}) + [E(\hat{a} - a)]^2\end{aligned}$$

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.

$$\begin{aligned}
\text{MSE}(\hat{a}) &= E[(\hat{a} - a)^2] \\
&= V(\hat{a}) + [E(\hat{a} - a)]^2 \\
&= \text{Variance}(\hat{a}) + \text{bias}(\hat{a})^2
\end{aligned}$$

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.

$$\begin{aligned}\text{MSE}(\hat{a}) &= E[(\hat{a}-a)^2] \\ &= V(\hat{a}) + [E(\hat{a}-a)]^2 \\ &= \text{Variance}(\hat{a}) + \text{bias}(\hat{a})^2\end{aligned}$$

To compare, compute

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.

$$\begin{aligned} \text{MSE}(\hat{a}) &= E[(\hat{a}-a)^2] \\ &= V(\hat{a}) + [E(\hat{a}-a)]^2 \\ &= \text{Variance}(\hat{a}) + \text{bias}(\hat{a})^2 \end{aligned}$$

To compare, compute

$$\text{M}SE(b_1^L) - \text{MSE}(b_1^O) =$$

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.

$$\begin{aligned}\text{MSE}(\hat{a}) &= E[(\hat{a}-a)^2]\\ &= V(\hat{a}) + [E(\hat{a}-a)]^2\\ &= \text{Variance}(\hat{a}) + \text{bias}(\hat{a})^2\end{aligned}$$

To compare, compute

$$\text{M}SE(b_1^L) - \text{MSE}(b_1^O) = \begin{cases} > 0 & \text{when omitting the variable is better} \\ < 0 & \text{when listwise deletion is better}\end{cases}$$

Derivation and Implications

Derivation and Implications

$$\text{MSE}(b_1^L) - \text{MSE}(b_1^O)$$

Derivation and Implications

$$\text{MSE}(b_1^L) - \text{MSE}(b_1^O) = \left(\frac{\lambda}{1-\lambda}\text{V}(b_1^I)\right) + F[\text{V}(b_2^I) - \beta_2\beta_2']F'$$

Derivation and Implications

$$\begin{aligned}\mathsf{MSE}(b_1^L) - \mathsf{MSE}(b_1^O) &= \left(\frac{\lambda}{1-\lambda}\mathsf{V}(b_1^I)\right) + F[\mathsf{V}(b_2^I) - \beta_2\beta_2']F' \\ &= (\text{Missingness part}) + (\text{Observed part})\end{aligned}$$

Derivation and Implications

$$\begin{aligned}\mathsf{MSE}(b_1^L) - \mathsf{MSE}(b_1^O) &= \left(\frac{\lambda}{1-\lambda}\mathsf{V}(b_1^I)\right) + F[\mathsf{V}(b_2^I) - \beta_2\beta_2']F' \\ &= (\text{Missingness part}) + (\text{Observed part})\end{aligned}$$

1. Missingness part (> 0) is an extra tilt away from listwise deletion

Derivation and Implications

$$\begin{aligned}\text{MSE}(b_1^L) - \text{MSE}(b_1^O) &= \left(\frac{\lambda}{1-\lambda}\text{V}(b_1^I)\right) + F[\text{V}(b_2^I) - \beta_2\beta_2']F' \\ &= (\text{Missingness part}) + (\text{Observed part})\end{aligned}$$

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

$$\begin{aligned}\text{MSE}(b_1^L) - \text{MSE}(b_1^O) &= \left(\frac{\lambda}{1-\lambda}\text{V}(b_1^I)\right) + F[\text{V}(b_2^I) - \beta_2\beta_2']F' \\ &= (\text{Missingness part}) + (\text{Observed part})\end{aligned}$$

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 λ usually? (from literature review in King et al 2001)

Derivation and Implications

$$\begin{aligned}\text{MSE}(b_1^L) - \text{MSE}(b_1^O) &= \left(\frac{\lambda}{1-\lambda}\text{V}(b_1^I)\right) + F[\text{V}(b_2^I) - \beta_2\beta_2']F' \\ &= (\text{Missingness part}) + (\text{Observed part})\end{aligned}$$

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 λ usually? (from literature review in King et al 2001)
    - $\lambda \approx 1/3$ on average in real political science articles

Derivation and Implications

$$\begin{aligned}\mathrm{MSE}(b_1^L) - \mathrm{MSE}(b_1^O) &= \left(\frac{\lambda}{1-\lambda}\mathrm{V}(b_1^I)\right) + F[\mathrm{V}(b_2^I) - \beta_2\beta_2']F' \\ &= \text{(Missingness part)} + \text{(Observed part)}\end{aligned}$$

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 λ 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

$$\begin{aligned}\text{MSE}(b_1^L) - \text{MSE}(b_1^O) &= \left(\frac{\lambda}{1-\lambda}\text{V}(b_1^I)\right) + F[\text{V}(b_2^I) - \beta_2\beta_2']F' \\ &= \text{(Missingness part)} + \text{(Observed part)}\end{aligned}$$

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 λ 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

4. If $\lambda \approx 0.5$, the contribution of the missingness (tilting away from choosing listwise deletion over omitting variables) is

Derivation and Implications

4. If $\lambda \approx 0.5$, the contribution of the missingness (tilting away from choosing listwise deletion over omitting variables) is

$$\text{RMSE difference} = \sqrt{\frac{\lambda}{1-\lambda} V(b_1^I)} = \sqrt{\frac{0.5}{1-0.5}} \text{SE}(b_1^I) = \text{SE}(b_1^I)$$

Derivation and Implications

4. If $\lambda \approx 0.5$, the contribution of the missingness (tilting away from choosing listwise deletion over omitting variables) is

$$\text{RMSE difference} = \sqrt{\frac{\lambda}{1-\lambda} V(b_1^I)} = \sqrt{\frac{0.5}{1-0.5}} \text{SE}(b_1^I) = \text{SE}(b_1^I)$$

(The sqrt of only one piece, for simplicity, not the difference.)

Derivation and Implications

4. If $\lambda \approx 0.5$, the contribution of the missingness (tilting away from choosing listwise deletion over omitting variables) is

$$\text{RMSE difference} = \sqrt{\frac{\lambda}{1-\lambda} V(b_1^I)} = \sqrt{\frac{0.5}{1-0.5}}\text{SE}(b_1^I) = \text{SE}(b_1^I)$$

   (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).

Derivation and Implications

4. If $\lambda \approx 0.5$, the contribution of the missingness (tilting away from choosing listwise deletion over omitting variables) is

$$\text{RMSE difference} = \sqrt{\frac{\lambda}{1-\lambda} V(b_1^I)} = \sqrt{\frac{0.5}{1-0.5}}\text{SE}(b_1^I) = \text{SE}(b_1^I)$$

(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 scenerio (MCAR)

The Best Case for Listwise Deletion

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 counter balance the biases induced by the other conditions.

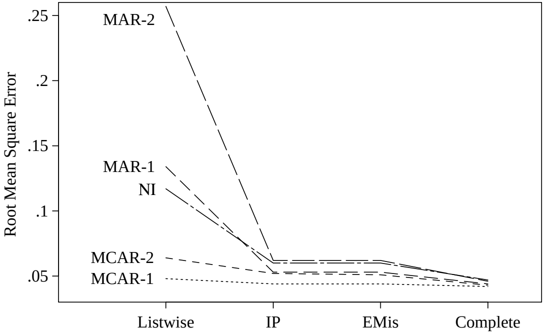
The Best Case for Listwise Deletion

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

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

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 propsective 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 propsective evaluations of national economic performance and personal financial circumstances
6. Control variables: age, education, family income, race, gender, union membership, ideology

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 propsective 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.

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 propsective 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.

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):

Listwise deletion	.43
	(.90)
Multiple imputation	1.65
	(.72)

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	1.65
	(.72)

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	.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	.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	.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

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):

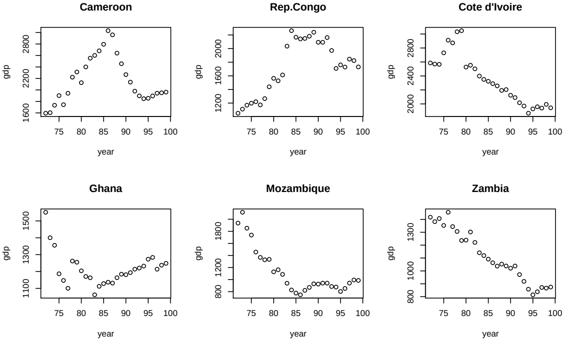
Listwise deletion	.43
	(.90)
Multiple imputation	1.65
	(.72)

so $(5 - 1) \times 1.65 = 6.6$, which is also a percentage of the range of Y.

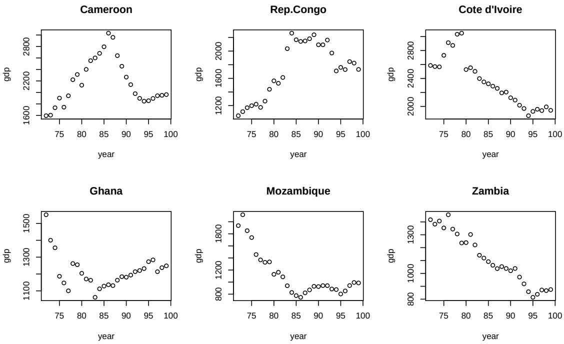
(a) MI estimator is more efficient, with a smaller SE
(b) The MI estimator is 4 times larger
(c) Based on listwise deletion, there is no evidence that perception of poor economic performance is related to support for Perot
(d) Based on the MI estimator, R's with negative retrospective economic evaluations are more likely to have favorable views of Perot.

MI in Time Series Cross-Section Data

MI in Time Series Cross-Section Data

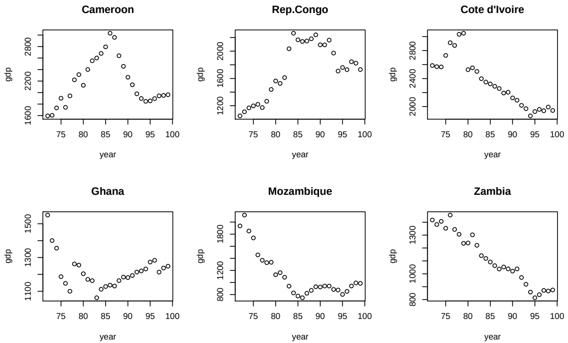


MI in Time Series Cross-Section Data



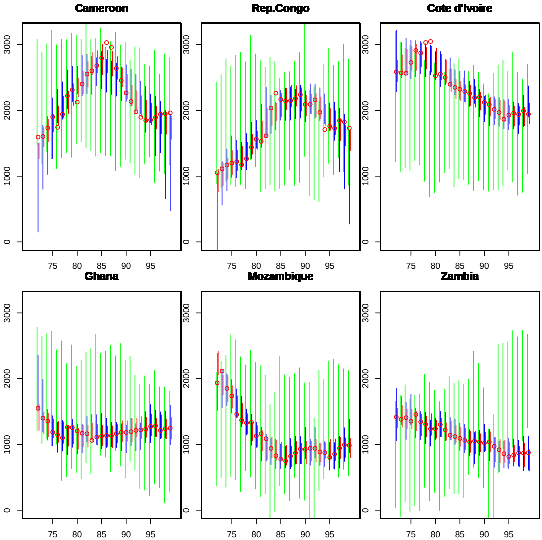
Include: (1) fixed effects, (2) time trends, and (3) priors for cells

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

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: $\ln p(\theta|y) = \ln[p(\theta)] + \sum_{i=1}^{n} \ln L_i(\theta|y)$

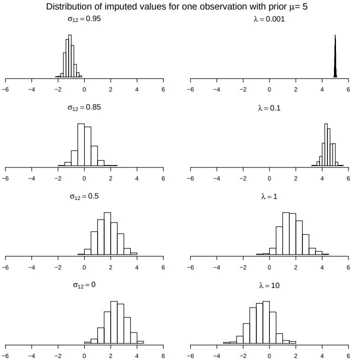
Priors on Cell Values

- 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

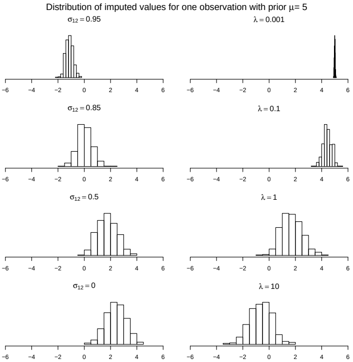
- Recall: $p(\theta|y) = p(\theta)\prod_{i=1}^{n} L_i(\theta|y)$
- take logs: $\ln p(\theta|y) = \ln[p(\theta)] + \sum_{i=1}^{n} \ln L_i(\theta|y)$
- $\implies$ Suppose prior is of the same form, $p(\theta|y) = L_i(\theta|y)$; then its just another observation: $\ln p(\theta|y) = \sum_{i=1}^{n+1} \ln L_i(\theta|y)$
- Honaker and King show how to modify these "data augmentation priors" to put priors on missing values rather than on μ and σ (or β).

Posterior imputation: mean=0, prior mean=5



Left column: holds prior $N(5, \lambda)$ constant ($\lambda = 1$) and changes predictive strength (the covariance, σ_{12}).

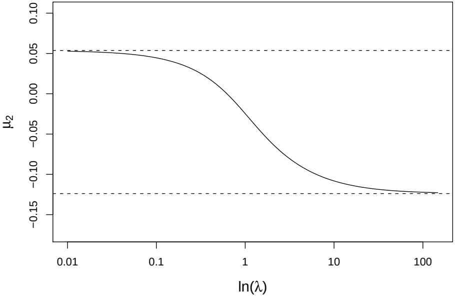
Posterior imputation: mean=0, prior mean=5



Left column: holds prior $N(5, \lambda)$ constant ($\lambda = 1$) and changes predictive strength (the covariance, σ_{12}).

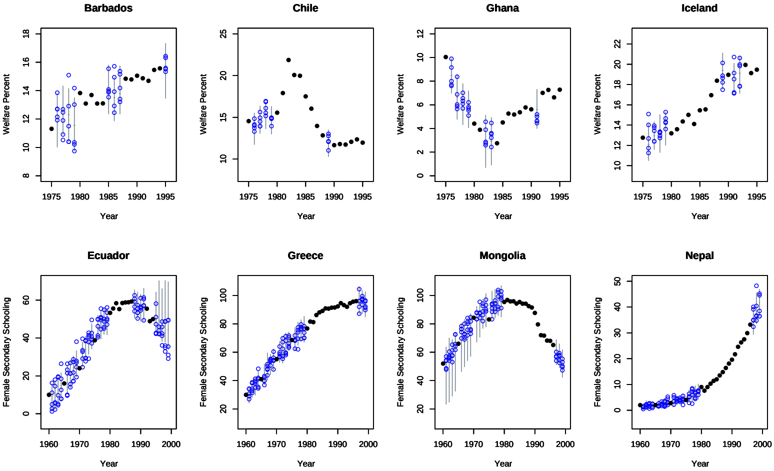
Right column: holds predictive strength of data constant (at $\sigma_{12} = 0.5$) and changes the strength of the prior (λ).

Model Parameters Respond to Prior on a Cell Value



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



Black = observed. Blue circles = five imputations; Bars = 95% CIs

	Listwise Deletion	Multiple Imputation
Life Expectancy		
Rich Democracies	−.072	.233
	(.179)	(.037)
Poor Democracies	−.082	.120
	(.040)	(.099)
N	1789	5627
Secondary Education		
Rich Democracies	.948	.948
	(.002)	(.019)
Poor Democracies	.373	.393
	(.094)	(.081)
N	1966	5627

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).