MATH 637: Mathematical Techniques in Data Science
The EM algorithm

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

May 13, 2020

Missing values in data

Missing data is a common problem in statistics.

- No measurement for a given individual/time/location, etc.
- Device failed.
- Error in data entry.
- Data was not disclosed for privacy reasons.
- etc.

Saundercock, Mr. William Henry	male	20.0	0
Andersson, Mr. Anders Johan	male	39.0	1
Vestrom, Miss. Hulda Amanda Adolfina	female	14.0	0
Hewlett, Mrs. (Mary D Kingcome)	female	55.0	0
Rice, Master. Eugene	male	2.0	4
Williams, Mr. Charles Eugene	male	NaN	0
Vander Planke, Mrs. Julius (Emelia Maria Vande...	female	31.0	1
Masselmani, Mrs. Fatima	female	NaN	0

Missing values in the titanic passengers dataset.

Missing values in data

Missing data is a common problem in statistics.

- No measurement for a given individual/time/location, etc.
- Device failed.
- Error in data entry.
- Data was not disclosed for privacy reasons.
- etc.

Saundercock, Mr. William Henry	male	20.0	0
Andersson, Mr. Anders Johan	male	39.0	1
Vestrom, Miss. Hulda Amanda Adolfina	female	14.0	0
Hewlett, Mrs. (Mary D Kingcome)	female	55.0	0
Rice, Master. Eugene	male	2.0	4
Williams, Mr. Charles Eugene	male	NaN	0
Vander Planke, Mrs. Julius (Emelia Maria Vande...	female	31.0	1
Masselmani, Mrs. Fatima	female	NaN	0

Missing values in the titanic passengers dataset.

How can we deal with missing values?

- Many possible procedures.
- The choice of the procedure can significantly impact the conclusions of a study.

Some strategies for dealing with missing values

Some options for dealing with missing values:

- **Deletion** (delete observations, remove variable, etc.). Solves the problem, but ignores some of the data (can be significant). May lead to ignoring an entire "category" of observations. Can generate significant bias.

Some strategies for dealing with missing values

Some options for dealing with missing values:

- **Deletion** (delete observations, remove variable, etc.).
  Solves the problem, but ignores some of the data (can be significant). May lead to ignoring an entire "category" of observations. Can generate significant bias.
- **Interpolation.**
  Sometimes it is possible to interpolate missing values (e.g. timeseries). However, we need enough data to be able to produce a good interpolation. In some problems, interpolation is not an option (e.g. age in the titanic passenger data).

Some strategies for dealing with missing values

Some options for dealing with missing values:

- **Deletion** (delete observations, remove variable, etc.).
  Solves the problem, but ignores some of the data (can be significant). May lead to ignoring an entire "category" of observations. Can generate significant bias.
- **Interpolation.**
  Sometimes it is possible to interpolate missing values (e.g. timeseries). However, we need enough data to be able to produce a good interpolation. In some problems, interpolation is not an option (e.g. age in the titanic passenger data).
- **Replace missing value with mean.**
  May introduce bias. Only valid for numerical observations.

Some strategies for dealing with missing values

Some options for dealing with missing values:

- **Deletion** (delete observations, remove variable, etc.).
  Solves the problem, but ignores some of the data (can be significant). May lead to ignoring an entire "category" of observations. Can generate significant bias.
- **Interpolation.**
  Sometimes it is possible to interpolate missing values (e.g. timeseries). However, we need enough data to be able to produce a good interpolation. In some problems, interpolation is not an option (e.g. age in the titanic passenger data).
- **Replace missing value with mean.**
  May introduce bias. Only valid for numerical observations.
- **Imputation with the EM algorithm.**
  Replace missing values by the *most likely values*. Account for all information available. Much more rigorous. However, requires a model. Can be computationally intensive.

Missing data mechanism

"Types" of missing data:

1. **Missing completely at random** (MCAR): The events that lead to a missing value are independent of both the *observable variables* and of the *unobservable parameters* of interest, and occur entirely at random. (Rarely the case in practice.)

Missing data mechanism

"Types" of missing data:

1. **Missing completely at random** (MCAR): The events that lead to a missing value are independent of both the *observable variables* and of the *unobservable parameters* of interest, and occur entirely at random. (Rarely the case in practice.)
2. **Missing at random** (MAR): missingness is not random, but can be fully accounted for by *observed values*.

Missing data mechanism

"Types" of missing data:

1. **Missing completely at random** (MCAR): The events that lead to a missing value are independent of both the *observable variables* and of the *unobservable parameters* of interest, and occur entirely at random. (Rarely the case in practice.)
2. **Missing at random** (MAR): missingness is not random, but can be fully accounted for by *observed values*.
3. **Missing not at random** (MNAR): neither MAR nor MCAR.

Missing data mechanism

"Types" of missing data:

1. **Missing completely at random** (MCAR): The events that lead to a missing value are independent of both the *observable variables* and of the *unobservable parameters* of interest, and occur entirely at random. (Rarely the case in practice.)
2. **Missing at random** (MAR): missingness is not random, but can be fully accounted for by *observed values*.
3. **Missing not at random** (MNAR): neither MAR nor MCAR.

Example: a study about people's weight. We measure (weight, sex).

Missing data mechanism

"Types" of missing data:

1. **Missing completely at random** (MCAR): The events that lead to a missing value are independent of both the *observable variables* and of the *unobservable parameters* of interest, and occur entirely at random. (Rarely the case in practice.)
2. **Missing at random** (MAR): missingness is not random, but can be fully accounted for by *observed values*.
3. **Missing not at random** (MNAR): neither MAR nor MCAR.

Example: a study about people's weight. We measure (weight, sex).

- Some respondent may not answer the survey for no particular reason. MCAR

Missing data mechanism

"Types" of missing data:

1. **Missing completely at random** (MCAR): The events that lead to a missing value are independent of both the *observable variables* and of the *unobservable parameters* of interest, and occur entirely at random. (Rarely the case in practice.)
2. **Missing at random** (MAR): missingness is not random, but can be fully accounted for by *observed values*.
3. **Missing not at random** (MNAR): neither MAR nor MCAR.

Example: a study about people's weight. We measure (weight, sex).

- Some respondent may not answer the survey for no particular reason. MCAR
- Maybe women are less likely to answer than male (independently of their weight). MAR

Missing data mechanism

"Types" of missing data:

1. **Missing completely at random** (MCAR): The events that lead to a missing value are independent of both the *observable variables* and of the *unobservable parameters* of interest, and occur entirely at random. (Rarely the case in practice.)
2. **Missing at random** (MAR): missingness is not random, but can be fully accounted for by *observed values*.
3. **Missing not at random** (MNAR): neither MAR nor MCAR.

Example: a study about people's weight. We measure (weight, sex).

- Some respondent may not answer the survey for no particular reason. MCAR
- Maybe women are less likely to answer than male (independently of their weight). MAR
- Heavy or light people may be less likely to disclose their weight. MNAR.

Example

- Suppose we have **independent** observations of a *discrete* random vector $X = (X_1, X_2, X_3, X_4)$ taking values in $\{0, 1, 2, 3\}$.

Example

- Suppose we have **independent** observations of a *discrete* random vector $X = (X_1, X_2, X_3, X_4)$ taking values in $\{0, 1, 2, 3\}$.

X_1	X_2	X_3	X_4
2	0	2	3
3	NA	1	1
1	3	NA	NA
2	NA	1	NA

Example

- Suppose we have **independent** observations of a *discrete* random vector $X = (X_1, X_2, X_3, X_4)$ taking values in $\{0, 1, 2, 3\}$.

X_1	X_2	X_3	X_4
2	0	2	3
3	NA	1	1
1	3	NA	NA
2	NA	1	NA

- Let $p(x_1, x_2, x_3, x_4) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$ be the pmf of X.

Example

- Suppose we have **independent** observations of a *discrete* random vector $X = (X_1, X_2, X_3, X_4)$ taking values in $\{0, 1, 2, 3\}$.

X_1	X_2	X_3	X_4
2	0	2	3
3	NA	1	1
1	3	NA	NA
2	NA	1	NA

- Let $p(x_1, x_2, x_3, x_4) = P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4)$ be the pmf of X.
- Ignoring the missing data mechanism, we have

$$p(x_1, \mathrm{NA}, x_3, x_4) = \sum_{x=0}^{3} p(x_1, x, x_2, x_3).$$

Example (cont.)

- Suppose the data comes from a parametric model $p(x_1, x_2, x_3, x_4; \theta)$ where $\theta \in \Theta$ is unknown.

X_1	X_2	X_3	X_4
2	0	2	3
3	NA	1	1
1	3	NA	NA
2	NA	1	NA

Example (cont.)

- Suppose the data comes from a parametric model $p(x_1, x_2, x_3, x_4; \theta)$ where $\theta \in \Theta$ is unknown.

X_1	X_2	X_3	X_4
2	0	2	3
3	NA	1	1
1	3	NA	NA
2	NA	1	NA

- We compute the *likelihood* of the data:

$$L(\theta) = p(2,0,2,3) \times p_{1,3,4}(3,1,1) \times p_{1,2}(1,3) \times p_{1,3}(2,1),$$

where $p_{1,3,4}(x_1, x_3, x_4) = \sum_{x_2=0}^{3} p(x_1, x_2, x_3, x_4)$, $p_{1,2}(x_1, x_2) = \sum_{x_3=0}^{3} \sum_{x_4=0}^{3} p(x_1, x_2, x_3, x_4)$, and $p_{1,3}(x_1, x_3) = \sum_{x_2=0}^{3} \sum_{x_4=0}^{3} p(x_1, x_2, x_3, x_4)$ denote *marginals* of p.

Example (cont.)

- Suppose the data comes from a parametric model $p(x_1, x_2, x_3, x_4; \theta)$ where $\theta \in \Theta$ is unknown.

X_1	X_2	X_3	X_4
2	0	2	3
3	NA	1	1
1	3	NA	NA
2	NA	1	NA

- We compute the *likelihood* of the data:

$$L(\theta) = p(2,0,2,3) \times p_{1,3,4}(3,1,1) \times p_{1,2}(1,3) \times p_{1,3}(2,1),$$

  where $p_{1,3,4}(x_1, x_3, x_4) = \sum_{x_2=0}^{3} p(x_1, x_2, x_3, x_4)$, $p_{1,2}(x_1, x_2) = \sum_{x_3=0}^{3} \sum_{x_4=0}^{3} p(x_1, x_2, x_3, x_4)$, and $p_{1,3}(x_1, x_3) = \sum_{x_2=0}^{3} \sum_{x_4=0}^{3} p(x_1, x_2, x_3, x_4)$ denote *marginals* of p.

- The *likelihood* can now be maximized as a function of θ.

Imputing the missing values

- Recall that $f(x) = E(Y|X = x)$ has the following optimality property:

$$E(Y|X = x) = \underset{c \in \mathbb{R}}{\operatorname{argmin}}\, E(Y - c)^2$$

where c is some function of x.

Imputing the missing values

- Recall that $f(x) = E(Y|X = x)$ has the following optimality property:

$$E(Y|X = x) = \underset{c \in \mathbb{R}}{\operatorname{argmin}} \, E(Y - c)^2$$

where c is some function of x.

- So $E(Y|X = x)$ is the "best prediction" of Y given X in the mean squared error sense.

Imputing the missing values

- Recall that $f(x) = E(Y|X = x)$ has the following optimality property:

$$E(Y|X = x) = \underset{c \in \mathbb{R}}{\operatorname{argmin}}\, E(Y - c)^2$$

where c is some function of x.

- So $E(Y|X = x)$ is the "best prediction" of Y given X in the mean squared error sense.
- As a result, once $p(x; \theta)$ is known (after estimating θ by maximum likelihood for example), we can *impute* missing values using:

$$\hat{x}_{\text{miss}} = E(x_{\text{miss}} | x_{\text{observed}}).$$

Imputing the missing values

- Recall that $f(x) = E(Y|X = x)$ has the following optimality property:

$$E(Y|X = x) = \underset{c \in \mathbb{R}}{\operatorname{argmin}}\, E(Y - c)^2$$

where c is some function of x.

- So $E(Y|X = x)$ is the "best prediction" of Y given X in the mean squared error sense.
- As a result, once $p(x; \theta)$ is known (after estimating θ by maximum likelihood for example), we can *impute* missing values using:

$$\hat{x}_{\text{miss}} = E(x_{\text{miss}} | x_{\text{observed}}).$$

For example, if $x = (1, 3, \text{NA}, \text{NA})$ then:

$$(\hat{x}_3, \hat{x}_4) = E((X_3, X_4)|X_1 = 1, X_2 = 3),$$

where E is computed with respect to $p(x_1, x_2, x_3, x_4; \theta)$.

Summary

In summary, given a family of probability models $p(x;\theta)$ for the data, under MAR, we can:

In summary, given a family of probability models $p(x;\theta)$ for the data, under MAR, we can:

1. Compute the likelihood of θ by *marginalizing* over the missing values.

Summary

In summary, given a family of probability models $p(x;\theta)$ for the data, under MAR, we can:

1. Compute the likelihood of θ by *marginalizing* over the missing values.
2. Estimate the parameter θ by maximum likelihood.

Summary

In summary, given a family of probability models $p(x;\theta)$ for the data, under MAR, we can:

1. Compute the likelihood of θ by *marginalizing* over the missing values.
2. Estimate the parameter θ by maximum likelihood.
3. Impute missing values using $\hat{x}_{\text{miss}} = E_\theta(x_{\text{miss}}|x_{\text{observed}})$, where E_θ denotes the expected value with respect to the probability distribution p_θ.

Summary

In summary, given a family of probability models $p(x;\theta)$ for the data, under MAR, we can:

1. Compute the likelihood of θ by *marginalizing* over the missing values.
2. Estimate the parameter θ by maximum likelihood.
3. Impute missing values using $\hat{x}_{\text{miss}} = E_\theta(x_{\text{miss}}|x_{\text{observed}})$, where E_θ denotes the expected value with respect to the probability distribution p_θ.

Remark: We assumed above that the variables are discrete, and the observations are independent for simplicity. The same procedure applied in the general case.

Problems

- The methodology described so far solves our missing data problem in principle.

Problems

- The methodology described so far solves our missing data problem in principle.
- However, explicitly finding the maximum of the likelihood function can be very difficult.

- The methodology described so far solves our missing data problem in principle.
- However, explicitly finding the maximum of the likelihood function can be very difficult.

The **Expectation-Maximization** (EM) algorithm of *Dempster, Laird, and Rubin, 1977* provides a more efficient way of solving the problem.

Problems

- The methodology described so far solves our missing data problem in principle.
- However, explicitly finding the maximum of the likelihood function can be very difficult.

The **Expectation-Maximization** (EM) algorithm of *Dempster, Laird, and Rubin, 1977* provides a more efficient way of solving the problem.

The EM algorithm leverages the fact the the likelihood is often easy to maximize if there is no missing values.

The EM algorithm

For simplicity, we will assume our observations are independent and the random variables are discrete.

For simplicity, we will assume our observations are independent and the random variables are discrete.

Some notation:

- We have a random vector W taking values in $\mathbb{R}^p$.

The EM algorithm

For simplicity, we will assume our observations are independent and the random variables are discrete.

Some notation:

- We have a random vector W taking values in $\mathbb{R}^p$.
- The distribution of the vector is $p(w; \theta)$.

The EM algorithm

For simplicity, we will assume our observations are independent and the random variables are discrete.

Some notation:

- We have a random vector W taking values in $\mathbb{R}^p$.
- The distribution of the vector is $p(w; \theta)$.
- We want to estimate θ.

The EM algorithm

For simplicity, we will assume our observations are independent and the random variables are discrete.

Some notation:

- We have a random vector W taking values in $\mathbb{R}^p$.
- The distribution of the vector is $p(w;\theta)$.
- We want to estimate θ.
- We only observe a part of the vector

$$(x^{(i)}, z^{(i)}) \in \mathbb{R}^{p_i} \times \mathbb{R}^{p-p_i} \qquad (i = 1, \ldots, n).$$

The EM algorithm

For simplicity, we will assume our observations are independent and the random variables are discrete.

Some notation:

- We have a random vector W taking values in $\mathbb{R}^p$.
- The distribution of the vector is $p(w;\theta)$.
- We want to estimate θ.
- We only observe a part of the vector

$$(x^{(i)}, z^{(i)}) \in \mathbb{R}^{p_i} \times \mathbb{R}^{p-p_i} \qquad (i = 1, \dots, n).$$

- So $x^{(i)}$ is the **observed** part and $z^{(i)}$ is the **unobserved** part.

The EM algorithm

For simplicity, we will assume our observations are independent and the random variables are discrete.

Some notation:

- We have a random vector W taking values in $\mathbb{R}^p$.
- The distribution of the vector is $p(w;\theta)$.
- We want to estimate θ.
- We only observe a part of the vector

$$(x^{(i)}, z^{(i)}) \in \mathbb{R}^{p_i} \times \mathbb{R}^{p-p_i} \qquad (i = 1, \dots, n).$$

- So $x^{(i)}$ is the **observed** part and $z^{(i)}$ is the **unobserved** part.
- The log-likelihood function is given by

$$l(\theta) = \sum_{i=1}^n \log p(x^{(i)};\theta) = \sum_{i=1}^n \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)};\theta).$$

(the second sum is over all the possible values of $z^{(i)}$).

The EM algorithm

For simplicity, we will assume our observations are independent and the random variables are discrete.

Some notation:

- We have a random vector W taking values in $\mathbb{R}^p$.
- The distribution of the vector is $p(w;\theta)$.
- We want to estimate θ.
- We only observe a part of the vector

$$(x^{(i)}, z^{(i)}) \in \mathbb{R}^{p_i} \times \mathbb{R}^{p-p_i} \qquad (i = 1, \dots, n).$$

- So $x^{(i)}$ is the **observed** part and $z^{(i)}$ is the **unobserved** part.
- The log-likelihood function is given by

$$l(\theta) = \sum_{i=1}^{n} \log p(x^{(i)};\theta) = \sum_{i=1}^{n} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)};\theta).$$

  (the second sum is over all the possible values of $z^{(i)}$).
- We would like to maximize that function over θ (generally difficult).

The EM algorithm (cont.)

Instead of trying to maximize the log-likelihood directly, the EM algorithm constructs a sequence of approximations $\theta^{(i)}$ of θ.

The EM algorithm (cont.)

Instead of trying to maximize the log-likelihood directly, the EM algorithm constructs a sequence of approximations $\theta^{(i)}$ of θ.

- Let $\theta^{(0)}$ be an **initial guess** for θ.

The EM algorithm (cont.)

Instead of trying to maximize the log-likelihood directly, the EM algorithm constructs a sequence of approximations $\theta^{(i)}$ of θ.

- Let $\theta^{(0)}$ be an **initial guess** for θ.
- Given the current estimate $\theta^{(i)}$ of θ, compute

$$\begin{aligned} Q(\theta|\theta^{(i)}) &:= E_{z|x;\theta^{(i)}} \ \log p(x,z;\theta) \\ &= \sum_{i=1}^{n} E_{z^{(i)}|x^{(i)};\theta^{(i)}} \left(\log p(x^{(i)}, z^{(i)}; \theta)\right) \qquad \text{(E step)} \end{aligned}$$

The EM algorithm (cont.)

Instead of trying to maximize the log-likelihood directly, the EM algorithm constructs a sequence of approximations $\theta^{(i)}$ of θ.

- Let $\theta^{(0)}$ be an **initial guess** for θ.
- Given the current estimate $\theta^{(i)}$ of θ, compute

$$\begin{aligned} Q(\theta|\theta^{(i)}) &:= E_{z|x;\theta^{(i)}} \ \log p(x, z; \theta) \\ &= \sum_{i=1}^{n} E_{z^{(i)}|x^{(i)};\theta^{(i)}} \left(\log p(x^{(i)}, z^{(i)}; \theta) \right) \qquad \text{(E step)} \end{aligned}$$

(In other words, we average the missing values according to their distribution after observing the observed values.)

The EM algorithm (cont.)

Instead of trying to maximize the log-likelihood directly, the EM algorithm constructs a sequence of approximations $\theta^{(i)}$ of θ.

- Let $\theta^{(0)}$ be an **initial guess** for θ.
- Given the current estimate $\theta^{(i)}$ of θ, compute

$$\begin{aligned} Q(\theta|\theta^{(i)}) &:= E_{z|x;\theta^{(i)}} \ \log p(x, z; \theta) \\ &= \sum_{i=1}^{n} E_{z^{(i)}|x^{(i)};\theta^{(i)}} \left(\log p(x^{(i)}, z^{(i)}; \theta)\right) \qquad \text{(E step)} \end{aligned}$$

(In other words, we average the missing values according to their distribution after observing the observed values.)

- We then optimize $Q(\theta|\theta^{(i)})$ with respect to θ:

$$\theta^{(i+1)} := \operatorname*{argmax}_{\theta} Q(\theta|\theta^{(i)}) \qquad \text{(M step)}.$$

We repeat this process until convergence.

The EM algorithm (cont.)

Theorem: The sequence $\theta^{(i)}$ constructed by the EM algorithm satisfies:

$$l(\theta^{(i+1)}) \geq l(\theta^{(i)}).$$

The EM algorithm (cont.)

Theorem: The sequence $\theta^{(i)}$ constructed by the EM algorithm satisfies:

$$l(\theta^{(i+1)}) \geq l(\theta^{(i)}).$$

- Hence, the sequence converges to a local max.

The EM algorithm (cont.)

Theorem: The sequence $\theta^{(i)}$ constructed by the EM algorithm satisfies:

$$l(\theta^{(i+1)}) \geq l(\theta^{(i)}).$$

- Hence, the sequence converges to a local max.
- With no additional assumptions, there is no guarantee that the EM algorithm will find the **global** max of the likelihood function.

The EM algorithm (cont.)

Theorem: The sequence $\theta^{(i)}$ constructed by the EM algorithm satisfies:

$$l(\theta^{(i+1)}) \geq l(\theta^{(i)}).$$

- Hence, the sequence converges to a local max.
- With no additional assumptions, there is no guarantee that the EM algorithm will find the **global** max of the likelihood function.
- Can use several starting points $\theta^{(0)}$ to increase the chances of finding a global maximum.

The EM algorithm (cont.)

Theorem: The sequence $\theta^{(i)}$ constructed by the EM algorithm satisfies:

$$l(\theta^{(i+1)}) \geq l(\theta^{(i)}).$$

- Hence, the sequence converges to a local max.
- With no additional assumptions, there is no guarantee that the EM algorithm will find the **global** max of the likelihood function.
- Can use several starting points $\theta^{(0)}$ to increase the chances of finding a global maximum.
- Once we reach convergence, we can estimate the missing values using the conditional expectation $\hat{x}_{\text{miss}} = E(x_{\text{miss}} | x_{\text{observed}}, \hat{\theta})$.