Lecture 3 gives the complete theory of the course at the abstract level.

The book strategy is to introduce the standard ad hoc techniques first—i.e., part I—before proceeding to the theory.
Key Idea:

There is data \( Y \), divided into two types:
- \((Y_{obs}, Y_{mis})\) — observable & missing data.

There are quantities/parameters \( \theta \) that we wish to estimate.

The data is generated by a model \( p(Y|\theta) \).

Case 1. If no data is missing, then we can estimate \( \theta \) by Maximum Likelihood (ML) or Maximum a Posteriori (MAP) — if we have prior \( P(\theta) \).

Case 2. If data is missing, we introduce an indicator variable \( M \) and a missing data mechanism \( P(M|Y, \phi) \) which may depend on unknown parameters \( \phi \).

This case will break down into subcases. Depending on whether the data is missing at random (by accident), or whether there is a bias. (E.G. height of army soldiers)
**Taxonomy of Missing Data Mechanisms**

- **MCAR** \( P(M \mid Y, \theta) = P(M \mid \theta) \) **Missing Completely at Random**
- **MAR** \( P(M \mid Y, \theta) = P(M \mid Y_{obs}, \theta) \) **Missing at Random**
- **NMAR** \( P(M \mid Y, \theta) \) **Not Missing at Random**

If no missing data, then MCAR.
\[ \hat{\theta} = \text{AKL} \max_{\theta} P(Y \mid \theta) \]

Otherwise
\[ P(M, Y \mid \theta, \phi) = P(M \mid Y, \phi) P(Y \mid \theta) \]
\[ P(M, Y_{obs} \mid \theta, \phi) = \int dY_{mis} P(M, Y_{obs}, Y_{mis} \mid \theta, \phi) \]

or \[
\sum_{Y_{mis}} P(M, Y_{obs}, Y_{mis} \mid \theta, \phi).
\]

You don't know \( \hat{\theta}, \hat{\phi} = \text{AKL} \max_{\theta, \phi} P(M, Y_{obs} \mid \theta, \phi) \)

**Special Cases:** if MCAR
\[ P(M, Y \mid \theta, \phi) = P(M \mid \theta) P(Y \mid \theta) \]
Then
\[ \hat{\phi} = \text{AKL} \max_{\phi} P(M \mid \theta) \]
\[ \hat{\theta} = \text{AKL} \max_{\theta} P(Y_{obs} \mid \theta) = \text{AKL} \max_{\theta} \sum_{Y_{mis}} P(Y_{obs}, Y_{mis} \mid \theta) \]

i.e. ML ignoring missing data.
\[ \text{if data is MAR} \quad P(M \mid Y_{\text{obs}}, \phi) = P(M \mid Y_{\text{obs}}, \phi) \]

\[ P(M, Y_{\text{obs}}, \text{Ymis} \mid \theta, \phi) = P(M \mid Y_{\text{obs}}, \phi) \cdot P(Y_{\text{obs}}, \text{Ymis} \mid \theta) \]

\[ P(M, Y_{\text{obs}} \mid \theta, \phi) = \sum_{\text{Ymis}} P(M, Y_{\text{obs}}, \text{Ymis} \mid \theta, \phi) \]

\[ = P(M \mid Y_{\text{obs}}, \phi) \cdot \sum_{\text{Ymis}} P(Y_{\text{obs}}, \text{Ymis} \mid \theta) \]

\[ = P(M \mid Y_{\text{obs}}, \phi) \cdot P(Y_{\text{obs}} \mid \theta). \]

Hence: \[ \hat{\theta} = \arg \max_{\theta} P(Y_{\text{obs}} \mid \theta) \] Can estimate \( \theta \)

\[ \hat{\phi} = \arg \max_{\phi} P(M \mid Y_{\text{obs}}, \phi) \] by ignoring the missing data.

But \text{if data is NMAR}

\[ P(M, Y_{\text{obs}}, \text{Ymis} \mid \theta, \phi) = P(M \mid Y_{\text{obs}}, \text{Ymis}, \phi) \cdot P(Y_{\text{obs}}, \text{Ymis} \mid \theta) \]

Hence:

\[ P(M, Y_{\text{obs}} \mid \theta, \phi) = \sum_{Y_{\text{obs}}} P(M \mid Y_{\text{obs}}, \text{Ymis}, \phi) \cdot P(Y_{\text{obs}} \mid \text{Ymis} \mid \theta) \]

Cannot be simplified.

In this situation, we cannot ignore the missing data mechanism. I.e., we will get a biased answer if we set \[ \hat{\theta} = \arg \max_{\theta} P(Y_{\text{obs}} \mid \theta) \]
MAP Estimation:

This introduces prior distributions \( P(\theta), P(\phi), \)
or even \( P(\theta, \phi) \) on the parameters (i.e. treats
them as random variables).

Priors are important if the number of
samples are small. Priors will have little
effect if the number of samples is large.

Priors encode knowledge about the
problem before we have obtained the
samples (prior = before Latin).

**MAP maximum a posteriori (MAP)**

Without missing data:

\[
\hat{\theta}_{MAP} = \arg \max_{\theta} \frac{P(y|\theta) P(\theta)}{P(y)} = \arg \max_{\theta} \frac{1}{P(y)} \prod_{i=1}^{n} P(y_i|\theta)
\]

\( \text{(Note } P(y) = \prod_{\theta} P(y|\theta) P(\theta) \text{)} \)

For large no. of samples:

\[
\hat{\theta}_{MAP} = \arg \max_{\theta} \sum_{i=1}^{n} \log P(y_i|\theta) + \log P(\theta)
\]

For large \( n \), \( \log P(y) \) is small.
MAP for NMAR case:

\( \hat{\theta}_{\text{MAP}} \), \( \hat{\phi}_{\text{MAP}} = \arg \max_{\theta, \phi} P(\mathbf{M}, \mathbf{Y}_{\text{obs}} | \theta, \phi) P(\theta, \phi) \)

Algorithms:

Algorithms are needed to perform ML and MAP estimation.

In some cases, (e.g. Gaussian distributions), the solution \( \hat{\theta} \) can be expressed as analytic functions of the data. (e.g. \( \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} Y_i \))

For NMAR, we must deal with the summation over the missing data \( Y_{\text{mis}} \).

There are two algorithms:

(i) Expectation Maximization (EM)
(ii) Data Augmentation (DA).

Both include ways to estimate (impute) the missing data \( Y_{\text{mis}} \).
The Height Example Revisited.

\[ P(y \mid \theta) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(y - \mu)^2}{2\sigma^2}} \quad \theta = (\mu, \sigma). \]

Suppose that the missing data mechanism is not deterministic.

\[ P(m = 0 \mid y, \phi) = \Phi(y - c; \phi) \quad \phi = (\sigma^2) \]
\[ P(m = 1 \mid y, \phi) = 1 - \Phi(y - c; \phi) \]

where \( \Phi(y - c; \phi) = \frac{1}{1 + e^{-(y - c)/\phi}}. \)

As \( \phi \rightarrow 0 \), this approaches the original sharp cut off at \( c \).

ie. \( P(m = 0 \mid y, \phi) = 1, \quad \text{if } y > c \)
\( = 0, \quad \text{if } y < c. \)

Joint distribution:

\[ P(m, y \mid \theta, \phi) = P(m \mid \phi, y) P(y \mid \theta) \]

Observed Data:
\[ y_i: \quad i = 1 \ldots N \]

Unobserved Data:
\[ \tilde{y}_i: \quad i = N + 1 \ldots M \]
\[ p(y_m, y_1, \Theta, \phi) = \prod_{i=1}^{N} p(m_i|\phi, y_i) p(y_i|\Theta) \]

\[ = \prod_{i=1}^{N} p(m_i|\phi, y_i) p(y_i|\Theta) \prod_{i=N+1}^{M} p(m_i|\phi, y_i) p(y_i|\Theta). \]

By definition:
\[ M_i = 0, \quad i = 1, \ldots, N \]
\[ M_i = 1, \quad i = N+1, \ldots, M. \]

\[ = \prod_{i=1}^{N} \bar{v}(y_i - c_i; 2) p(y_i|\Theta) \]
\[ \times \prod_{i=N+1}^{M} \{ 1 - \bar{v}(y_i - c_i; 2) \} p(y_i|\Theta). \]

**We need to integrate out w.r.t. \{ y_i : i = N+1 to M \}.**

Let \[ \psi(c, \gamma, \Theta) = \int_{-\infty}^{\infty} dy \{ 1 - \bar{v}(y - c; 2) \} p(y|\Theta) \]

Then \[ (\hat{\Theta}, \hat{\phi}) = \text{A.K.S. MAX} \prod_{i=1}^{N} \bar{v}(y_i - c_i; 2) p(y_i|\Theta) \]
\[ \prod_{i=N+1}^{M} \psi(c, \gamma, \Theta)^{M-N} \]

**Note:** That to do this means that we have to know how many samples were rejected (i.e. \( M \)).

If \( r = 0 \), \( \psi = 1 \), and we don't need to know \( M \).

Otherwise, put a prior on \( M \) and sum it out.
Overview:

1) This lecture included both ML, MAP formulation.

2) Summing out variables if you do not know them and do not need to estimate them. (Integrate out if variables are continuous)

3) Note: graphs can be used to represent the dependencies between variables.