BAYESIAN INFERENCE

Antoine Deleforge
THANKS

Prof. Dr.-Ing. Walter Kellermann
Head of LMS chair
EARS project coordinator

Dr. Roland Maas
PhD at LMS chair
Research and development at Amazon
• What is Bayesian inference?
  – Overview
  – Classical vs. Bayesian approach
  – Bayes Theorem & Example
  – General Methodology
• Bayesian inference by examples
  – Direct inference
  – The Expectation-Maximization algorithm
  – Variational Bayes methods
• Markov chain Monte Carlo
OUTLINE

• What is Bayesian inference?
  – Overview
  – Classical vs. Bayesian approach
  – Bayes Theorem & Example
  – General Methodology

• Bayesian inference by examples
  – Direct inference
  – The Expectation-Maximization algorithm
  – Variational Bayes methods

• Markov chain Monte Carlo
What is Bayesian Inference?

Inference
What is Bayesian Inference?

- **Ingredients**
  - Observations
  - Model

- **Inference**

---

Antoine.Deleforge@FAU.de  LVA/ICA Summer School 2015 5/42
What is Bayesian Inference?

Inference

**Ingredients**

- Observations
- Model

**Goals**

- **Estimation**
  
  Quantitative deductions on causes or consequences of the observations, i.e., find underlying model parameters.

  Example: *I observed a certain amount of rain drops forming on my window in the last minute. What is the current rainfall in millimeters?*
What is Bayesian Inference?

Inference

**Ingredients**
- Observations
- Model

**Goals**
- **Estimation**
  - Quantitative deductions on causes or consequences of the observations, i.e., find underlying model parameters.
  - Example: *I observed a certain amount of rain drops forming on my window in the last minute. What is the current rainfall in milimeters?*

- **Prediction**
  - From the inferred model, predict what missing or future observations should be.
  - Example: *How many more raindrops will form on my window in the next hour?*
What is Bayesian Inference?

Inference

Ingredients
- Observations
- Model

Goals
- **Estimation**: Quantitative deductions on causes or consequences of the observations, i.e., find underlying model parameters.
  
  Example: *I observed a certain amount of rain drops forming on my window in the last minute. What is the current rainfall in millimeters?*

- **Prediction**: From the inferred model, predict what missing or future observations should be.
  
  Example: *How many more raindrops will form on my window in the next hour?*

- **Decision**: Take a decision out of a discrete set of choices
  
  Example: *Is it safe to open my window 1 minute to get some fresh air?*
Statistics: Inference from the real world observations of a random phenomenon using probability theory
Statistics: Inference from the real world observations of a random phenomenon using probability theory

<table>
<thead>
<tr>
<th>Classical/Frequentist Statistics</th>
<th>Bayesian Statistics</th>
</tr>
</thead>
</table>

**Definition**

**Features**

**Tools**
Statistics: Inference from the real world observations of a random phenomenon using probability theory

<table>
<thead>
<tr>
<th>Classical/Frequentist Statistics</th>
<th>Bayesian Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>The model parameters which should be estimated are considered as unknown constant.</td>
<td></td>
</tr>
</tbody>
</table>
**Statistics**: Inference from the real world observations of a random phenomenon using probability theory

<table>
<thead>
<tr>
<th>Classical/Frequentist Statistics</th>
<th>Bayesian Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>The model parameters which should be estimated are considered as <strong>unknown constant</strong>.</td>
<td>The model parameters are considered as <strong>hidden random variables</strong>, following a hypothetical <strong>probabilistic model</strong>.</td>
</tr>
</tbody>
</table>
**Statistics:** Inference from the real world observations of a random phenomenon using probability theory

<table>
<thead>
<tr>
<th>Classical/Frequentist Statistics</th>
<th>Bayesian Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>The model parameters which should be estimated are considered as unknown constant.</td>
<td>The model parameters are considered as hidden random variables, following a hypothetical probabilistic model.</td>
</tr>
</tbody>
</table>

**Definition**

**Features**

Inference entirely based on observed data and frequentist arguments. Useful when few prior knowledge exist on the underlying random process.
**What is Bayesian Inference?**

**Statistics:** Inference from the real world observations of a random phenomenon using probability theory

<table>
<thead>
<tr>
<th>Classical/Frequentist Statistics</th>
<th>Bayesian Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>The model parameters which should be estimated are considered as <strong>unknown constant</strong>.</td>
<td>The model parameters are considered as <strong>hidden random variables</strong>, following a hypothetical <strong>probabilistic model</strong>.</td>
</tr>
<tr>
<td>Inference entirely based on <strong>observed data</strong> and <strong>frequentist arguments</strong>. Useful when few prior knowledge exist on the underlying random process.</td>
<td>Incorporate <strong>prior knowledge on the hidden variables</strong> in the form of a generative probabilistic model. Useful when some <strong>reasonable</strong> probability density functions (PDFs) can be assumed.</td>
</tr>
</tbody>
</table>
Statistics: Inference from the real world observations of a random phenomenon using probability theory

### Classical/Frequentist Statistics
- The model parameters which should be estimated are considered as **unknown constant**.
- Inference entirely based on **observed data** and **frequentist arguments**. Useful when few prior knowledge exist on the underlying random process.
- Tools:
  - Linear estimators
  - First and second order statistics
  - A lot of: \( \mathbb{E} \{ . . . \} \)

### Bayesian Statistics
- The model parameters are considered as **hidden random variables**, following a hypothetical **probabilistic model**.
- Incorporate **prior knowledge on the hidden variables** in the form of a generative probabilistic model. Useful when some **reasonnable** probability density functions (PDFs) can be assumed.

---

**Definition**

**Features**

**Tools**
**Statistics**: Inference from the real world observations of a random phenomenon using probability theory

<table>
<thead>
<tr>
<th>Classical/Frequentist Statistics</th>
<th>Bayesian Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>The model parameters which should be estimated are considered as <strong>unknown constant</strong>.</td>
<td>The model parameters are considered as <strong>hidden random variables</strong>, following a hypothetical <strong>probabilistic model</strong>.</td>
</tr>
<tr>
<td>Inference entirely based on <strong>observed data</strong> and <strong>frequentist arguments</strong>. Useful when few prior knowledge exist on the underlying random process.</td>
<td>Incorporate <strong>prior knowledge on the hidden variables</strong> in the form of a generative probabilistic model. Useful when some <strong>reasonable</strong> probability density functions (PDFs) can be assumed.</td>
</tr>
</tbody>
</table>

**Tools**

- Linear estimators
- First and second order statistics
- A lot of: \( \mathbb{E} \{ \ldots \} \)
- **Bayes’ Theorem**
- Explicit PDFs
- A lot of: \( \mathcal{N}(\ldots) \)
What is Bayesian Inference?

- Bayes’ Theorem

\[
p(Z = z | X = x) = \frac{p(X = x | Z = z)p(Z = z)}{p(X = x)}
\]
What is Bayesian Inference?

- Bayes’ Theorem

\[ p(Z = z \mid X = x) = \frac{p(X = x \mid Z = z)p(Z = z)}{p(X = x)} \]

\( X \): Observed variables  \( Z \): Hidden variables
• Bayes’ Theorem

\[
p(Z = z | X = x) = \frac{p(X = x | Z = z)p(Z = z)}{p(X = x)}
\]

- \(X\): Observed variables
- \(Z\): Hidden variables

Posterior
What is Bayesian Inference?

- **Bayes’ Theorem**

\[ p(Z = z | X = x) = \frac{p(X = x | Z = z) p(Z = z)}{p(X = x)} \]

\(X\) : Observed variables

\(Z\) : Hidden variables

Posterior

Likelihood
What is Bayesian Inference?

• Bayes’ Theorem

\[ p(Z = z | X = x) = \frac{p(X = x | Z = z)p(Z = z)}{p(X = x)} \]

\( X \): Observed variables

\( Z \): Hidden variables

Posterior

Likelihood

Prior
What is Bayesian Inference?

• Bayes’ Theorem

\[ p(Z = z | X = x) = \frac{p(X = x | Z = z)p(Z = z)}{p(X = x)} \]

- \( X \): Observed variables
- \( Z \): Hidden variables
- Posterior
- Likelihood
- Prior

« Observed data » or « marginal » likelihood
Bayes’ Theorem

\[ p(Z = z | X = x) = \frac{p(X = x | Z = z)p(Z = z)}{p(X = x)} \]

Remark 1: Bayes does not « forbid » model parameters!
- No formal difference between a parameter and a hidden variable with constant prior
- Priors distributions often have parameters called « hyperparameters »
## What is Bayesian Inference?

### Bayes’ Theorem

**Remark 1:** Bayes does not « forbid » model parameters!

- No formal difference between a parameter and a hidden variable with constant prior
- Priors distributions often have parameters called « hyperparameters »

\[
p(Z = z | X = x; \theta) = \frac{p(X = x | Z = z; \theta)p(Z = z; \theta)}{p(X = x; \theta)}
\]

\(X:\) Observed variables \(Z:\) Hidden variables

- Posterior
- Likelihood
- Prior

« Observed data » or « marginal » likelihood
What is Bayesian Inference?

• **Bayes’ Theorem**

\[
p(Z = z | X = x; \theta) = \frac{p(X = x | Z = z; \theta)p(Z = z; \theta)}{p(X = x; \theta)}
\]

- **Posterior**
- **Likelihood**
- **Prior**

**Remark 1:** Bayes does not « forbid » model parameters!
- No formal difference between a parameter and a hidden variable with constant prior
- Priors distributions often have parameters called « hyperparameters »

**Remark 2:** Why hidden variables?
- Formally not needed: \( p(X) \) can be obtained by marginalizing out hidden variables
- A convenient and powerful view point which makes inference possible in complex scenarios through a variety of methods
An Example:

Observations

<table>
<thead>
<tr>
<th>$X_D$</th>
<th>$X_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I see drops on my window</td>
<td>I see a cat or a dog at my window</td>
</tr>
</tbody>
</table>
An Example:

**Observations**

<table>
<thead>
<tr>
<th>$X_D$</th>
<th>$X_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I see drops on my window</td>
<td>I see a cat or a dog at my window</td>
</tr>
</tbody>
</table>

**Hypothesis**

<table>
<thead>
<tr>
<th>$Z = 0$</th>
<th>$Z = 1$</th>
<th>$Z = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>It’s not raining</td>
<td>It’s raining rain</td>
<td>It’s raining cats and dogs</td>
</tr>
</tbody>
</table>
An Example:

Observations

<table>
<thead>
<tr>
<th>$X_D$</th>
<th>$X_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I see drops on my window</td>
<td>I see a cat or a dog at my window</td>
</tr>
</tbody>
</table>

Hypothesis

<table>
<thead>
<tr>
<th>$Z$</th>
<th>$Z = 0$</th>
<th>$Z = 1$</th>
<th>$Z = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>It’s not raining</td>
<td>It’s raining rain</td>
<td>It’s raining cats and dogs</td>
</tr>
</tbody>
</table>

$p(X_i | Z = j)$

<table>
<thead>
<tr>
<th></th>
<th>$X_D$</th>
<th>$X_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z = 0$</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>$Z = 1$</td>
<td>0.99</td>
<td>0.1</td>
</tr>
<tr>
<td>$Z = 2$</td>
<td>0.1</td>
<td>0.99</td>
</tr>
</tbody>
</table>
An Example:

Observations

<table>
<thead>
<tr>
<th>$X_D$</th>
<th>$X_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I see drops on my window</td>
<td>I see a cat or a dog at my window</td>
</tr>
</tbody>
</table>

Hypothesis

<table>
<thead>
<tr>
<th>$Z = 0$</th>
<th>$Z = 1$</th>
<th>$Z = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>It’s not raining</td>
<td>It’s raining rain</td>
<td>It’s raining cats and dogs</td>
</tr>
</tbody>
</table>

Bayes’ Theorem & Example

\[
p(X_i | Z = j)\]

<table>
<thead>
<tr>
<th></th>
<th>$X_D$</th>
<th>$X_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z = 0$</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>$Z = 1$</td>
<td>0.99</td>
<td>0.1</td>
</tr>
<tr>
<td>$Z = 2$</td>
<td>0.1</td>
<td>0.99</td>
</tr>
</tbody>
</table>

\[
p(X_D, X_C | Z = 1) = p(X_D, X_C | Z = 2) = 0.099 : \text{equal likelihood!}\]
An Example:

Observations

<table>
<thead>
<tr>
<th>X_D</th>
<th>X_C</th>
</tr>
</thead>
<tbody>
<tr>
<td>I see drops on my window</td>
<td>I see a cat or a dog at my window</td>
</tr>
</tbody>
</table>

Hypothesis

<table>
<thead>
<tr>
<th>Z = 0</th>
<th>Z = 1</th>
<th>Z = 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>It’s not raining</td>
<td>It’s raining rain</td>
<td>It’s raining cats and dogs</td>
</tr>
</tbody>
</table>

\[ p(X_i | Z = j) \]

\[
\begin{array}{c|c|c}
  Z = 0 & Z = 1 & Z = 2 \\
  \hline
  X_D & 0.1 & 0.99 & 0.1 \\
  X_C & 0.1 & 0.1 & 0.99 \\
\end{array}
\]

\[ p(X_D, X_C | Z = 1) = p(X_D, X_C | Z = 2) = 0.099 : \text{equal likelihood!} \]

- **Add priors:**
  \[ p(Z = 0) = 49.995\% \quad p(Z = 1) = 49.995\% \quad p(Z = 2) = 0.01\% \]
What is Bayesian Inference?

An Example:

**Observations**

<table>
<thead>
<tr>
<th>$X_D$</th>
<th>$X_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I see drops on my</td>
<td>I see a cat or a dog</td>
</tr>
<tr>
<td>window</td>
<td>at my window</td>
</tr>
</tbody>
</table>

**Hypothesis**

<table>
<thead>
<tr>
<th>$Z = 0$</th>
<th>$Z = 1$</th>
<th>$Z = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>It’s not raining</td>
<td>It’s raining</td>
<td>It’s raining</td>
</tr>
<tr>
<td>rain</td>
<td>rain</td>
<td>cats and dogs</td>
</tr>
</tbody>
</table>

$p(X_i|Z = j)$

<table>
<thead>
<tr>
<th></th>
<th>$X_D$</th>
<th>$X_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z = 0$</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>$Z = 1$</td>
<td>0.99</td>
<td>0.1</td>
</tr>
<tr>
<td>$Z = 2$</td>
<td>0.1</td>
<td>0.99</td>
</tr>
</tbody>
</table>

$p(X_D, X_C|Z = 1) = p(X_D, X_C|Z = 2) = 0.099$: equal likelihood!

- **Add priors:** $p(Z = 0) = 49.995\%$  $p(Z = 1) = 49.995\%$  $p(Z = 2) = 0.01\%$

- **Bayes’ theorem:**

\[
p(Z = 1|X_D, X_C) \approx 99.98\% \quad p(Z = 2|X_D, X_C) \approx 0.02\%
\]
What is Bayesian Inference?

General Methodology
General Methodology

Modeling
General Methodology

What is hidden?

What is observed?

Dependencies?

Modeling
What is Bayesian Inference?

General Methodology

Modeling

- What is hidden?
- What is observed?
- Dependencies?

Graphical model
General Methodology

What is hidden?
What is observed?
Dependencies?

Graphical model
Choice of prior and conditional PDFs

Modeling
What is Bayesian Inference?

General Methodology

Modeling

- What is hidden?
- What is observed?
- Dependencies?

Graphical model + Choice of prior and conditional PDFs = Joint PDF
What is Bayesian Inference?

General Methodology

Modeling

What is hidden?
What is observed?
Dependencies?

Graphical model
Choice of prior and conditional PDFs

Inference
Joint PDF
What is Bayesian Inference?

General Methodology

Modeling

- What is hidden?
- What is observed?
- Dependencies?

Graphical model

Choice of prior and conditional PDFs

Joint PDF

Inference

Apply Baye’s Theorem
What is Bayesian Inference?

General Methodology

Modeling

- What is hidden?
- What is observed?
- Dependencies?

Graphical model

Choice of prior and conditional PDFs

Joint PDF

Inference

Apply Baye’s Theorem

Choice of method:
- Exact / Approximate
- Direct / Iterative
What is Bayesian Inference?

General Methodology

Modeling

- What is hidden?
- What is observed?
- Dependencies?

Inference

Apply Baye’s Theorem

Choice of method:
- Exact / Approximate
- Direct / Iterative

Graphical model + Choice of prior and conditional PDFs = Joint PDF

Posterior PDF
What is Bayesian Inference?

General Methodology

Modeling

- What is hidden?
- What is observed?
- Dependencies?

Graphical model

Choice of prior and conditional PDFs

Joint PDF

Inference

Apply Baye’s Theorem

Choice of method:
- Exact / Approximate
- Direct / Iterative

Posterior PDF

- Estimation (MAP, posterior mean, ...)
- Prediction
- Decision
• What is Bayesian inference?
  – Overview
  – Classical vs. Bayesian approach
  – Bayes Theorem & Example
  – General Methodology

• Bayesian inference by examples
  – Direct inference
  – The Expectation-Maximization algorithm
  – Variational Bayes methods

• Markov chain Monte Carlo
• What is Bayesian inference?
  – Overview
  – Classical vs. Bayesian approach
  – Bayes Theorem & Example
  – General Methodology

• Bayesian inference by examples
  – Direct inference
  – The Expectation-Maximization algorithm
  – Variational Bayes methods

• Markov chain Monte Carlo
• What is Bayesian inference?
  – Overview
  – Classical vs. Bayesian approach
  – Bayes Theorem & Example
  – General Methodology

• Bayesian inference by examples
  – Direct inference
  – The Expectation-Maximization algorithm
  – Variational Bayes methods

• Markov chain Monte Carlo
Bayesian Inference: Examples

Direct Inference
Bayesian Inference: Examples

Direct Inference
Bayesian Inference: Examples

Direct Inference
Bayesian Inference: Examples

Direct Inference
Bayesian Inference: Examples

Direct Inference
Bayesian Inference: Examples

Somemone is making a joke...
Bayesian Inference: Examples
**Bayesian Inference: Examples**

**Modeling**

**Observed variables:** \( \{x_n\}_{n=1}^{N} \subseteq \mathbb{R}^2 \)
Bayesian Inference: Examples

Modeling

Observed variables: \( \{x_n\}_{n=1}^{N} \subset \mathbb{R}^2 \)

Hidden variable: \( Z \in \{1, 2, 3\} \)

Guilty house number?
Bayesian Inference: Examples

Observed variables: $\{x_n\}_{n=1}^N \subset \mathbb{R}^2$

Hidden variable: $Z \in \{1, 2, 3\}$

Guilty house number?

Graphical Model:
Bayesian Inference: Examples

Modeling

Observed variables: \( \{x_n\}_{n=1}^{N} \subset \mathbb{R}^2 \)

Hidden variable: \( Z \in \{1, 2, 3\} \)

Graphical Model:

Conditionals

\[
\begin{align*}
p(X_n = x_n | Z = 1) &= \mathcal{N}(x_n; \mu_1, I) \\
p(X_n = x_n | Z = 2) &= \mathcal{N}(x_n; \mu_2, I) \\
p(X_n = x_n | Z = 3) &= \mathcal{N}(x_n; \mu_3, I)
\end{align*}
\]
Bayesian Inference: Examples

- Observed variables: \( \{x_n\}_{n=1}^{N} \subset \mathbb{R}^2 \)
- Hidden variable: \( Z \in \{1, 2, 3\} \)

Graphical Model:

Conditionals:

\[
\begin{align*}
    p(X_n = x_n | Z = 1) &= \mathcal{N}(x_n; \mu_1, I) \\
    p(X_n = x_n | Z = 2) &= \mathcal{N}(x_n; \mu_2, I) \\
    p(X_n = x_n | Z = 3) &= \mathcal{N}(x_n; \mu_3, I)
\end{align*}
\]
Bayesian Inference: Examples

Observed variables: \( \{x_n\}_{n=1}^{N} \subset \mathbb{R}^2 \)

Hidden variable: \( Z \in \{1, 2, 3\} \)

Graphical Model:

Conditionals:
\[
\begin{align*}
p(X_n = x_n | Z = 1) &= \mathcal{N}(x_n; \mu_1, I) \\
p(X_n = x_n | Z = 2) &= \mathcal{N}(x_n; \mu_2, I) \\
p(X_n = x_n | Z = 3) &= \mathcal{N}(x_n; \mu_3, I)
\end{align*}
\]

Priors:
\[
\begin{align*}
p(Z = 1) &= 0.1 \quad \text{(Grandma Jane)} \\
p(Z = 2) &= 0.3 \quad \text{(Student house)} \\
p(Z = 3) &= 0.6 \quad \text{(Family with kids)}
\end{align*}
\]
Bayesian Inference: Examples

Bayes’ Theorem:

\[ p(Z = i|\mathbf{x}_1, \ldots, \mathbf{x}_N) = \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_N|Z = i)p(Z = i)}{p(\mathbf{x}_1, \ldots, \mathbf{x}_N)} \]
Bayesian Inference: Examples

Bayes’ Theorem:

\[ p(Z = i | \mathbf{x}_1, \ldots, \mathbf{x}_N) = \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_N | Z = i) p(Z = i)}{p(\mathbf{x}_1, \ldots, \mathbf{x}_N)} \]

\[ = \frac{\prod_{n=1}^{N} p(\mathbf{x}_n | Z = i) p(Z = i)}{\sum_{k=1}^{3} \prod_{n=1}^{N} p(\mathbf{x}_n | Z = k) p(Z = k)} \]
Bayesian Inference: Examples

Bayes’ Theorem:

\[
p(Z = i | \mathbf{x}_1, \ldots, \mathbf{x}_N) = \frac{p(\mathbf{x}_1, \ldots, \mathbf{x}_N | Z = i)p(Z = i)}{p(\mathbf{x}_1, \ldots, \mathbf{x}_N)} = \frac{\prod_{n=1}^{N} p(\mathbf{x}_n | Z = i)p(Z = i)}{\sum_{k=1}^{N} \prod_{n=1}^{N} p(\mathbf{x}_n | Z = k)p(Z = k)}
\]

Direct computation
Bayesian Inference: Examples

Bayes’ Theorem:

\[
p(Z = i | x_1, \ldots, x_N) = \frac{p(x_1, \ldots, x_N | Z = i) p(Z = i)}{p(x_1, \ldots, x_N)}
\]

\[
= \frac{\prod_{n=1}^{N} p(x_n | Z = i) p(Z = i)}{\sum_{k=1}^{3} \prod_{n=1}^{N} p(x_n | Z = k) p(Z = k)}
\]

Estimation: Maximum a Posteriori (MAP)

\[
\hat{z} = \arg\max_i \left[ p(Z = i | x_1, \ldots, x_N) \right]
\]
Bayes’ Theorem:

\[
p(Z = i | x_1, \ldots, x_N) = \frac{p(x_1, \ldots, x_N | Z = i)p(Z = i)}{p(x_1, \ldots, x_N)}
\]

\[= \frac{\prod_{n=1}^{N} p(x_n | Z = i)p(Z = i)}{\sum_{k=1}^{3} \prod_{n=1}^{N} p(x_n | Z = k)p(Z = k)}\]

**Estimation:** Maximum a Posteriori (MAP)

\[
\hat{z} = \arg\max_i \left[ p(Z = i | x_1, \ldots, x_N) \right] \quad \Rightarrow \quad \hat{z} = 2, \text{ the student house}
\]
Bayes’ Theorem:

\[ p(Z = i | x_1, \ldots, x_N) = \frac{p(x_1, \ldots, x_N | Z = i)p(Z = i)}{p(x_1, \ldots, x_N)} = \frac{\prod_{n=1}^{N} p(x_n | Z = i)p(Z = i)}{\sum_{k=1}^{3} \prod_{n=1}^{N} p(x_n | Z = k)p(Z = k)} \]

Estimation: Maximum a Posteriori (MAP)

\[ \hat{z} = \arg\max_{i} p(Z = i | x_1, \ldots, x_N) \implies \hat{z} = 2, \text{ the student house} \]

Decision: These pranksters will hear from me at the Uni!
• What is Bayesian inference?
  – Overview
  – Classical vs. Bayesian approach
  – Bayes Theorem & Example
  – General Methodology

• Bayesian inference by examples
  – Direct inference
  – The Expectation-Maximization algorithm
  – Variational Bayes methods

• Markov chain Monte Carlo
OUTLINE

• What is Bayesian inference?
  – Overview
  – Classical vs. Bayesian approach
  – Bayes Theorem & Example
  – General Methodology

• Bayesian inference by examples
  – Direct inference
  – The Expectation-Maximization algorithm
  – Variational Bayes methods

• Markov chain Monte Carlo
Bayesian Inference: Examples
Bayesian Inference: Examples

EM algorithm
Bayesian Inference: Examples

EM algorithm
They are on the roof!
Bayesian Inference: Examples

Modeling
Bayesian Inference: Examples

Modeling

Observed variables: \( \{ x_n \}_{n=1}^N \subset \mathbb{R}^2 \)
Bayesian Inference: Examples

**Observed variables:** \( \{x_n\}_{n=1}^{N} \subset \mathbb{R}^2 \)

**Hidden Variables:** \( \{Z_n\}_{n=1}^{N} \in \{1, 2, 3\}^N \)

Modeling

1. 2. 3.

EM algorithm
Bayesian Inference: Examples

Observed variables: \( \{x_n\}_{n=1}^{N} \subset \mathbb{R}^2 \)

Graphical Model:

Hidden Variables: \( \{Z_n\}_{n=1}^{N} \in \{1, 2, 3\}^N \)
Bayesian Inference: Examples

Modeling

**Observed variables:** \( \{ x_n \}_{n=1}^N \subset \mathbb{R}^2 \)

**Hidden Variables:** \( \{ Z_n \}_{n=1}^N \in \{1, 2, 3\}^N \)

**Graphical Model:**

**Conditional:**

\[
p(X_n = x_n | Z_n = k; \theta) = \mathcal{N}(x_n; \mu_k, \Sigma_k)
\]
Bayesian Inference: Examples

**Observed variables:** \( \{x_n\}_{n=1}^N \subset \mathbb{R}^2 \)

**Hidden Variables:** \( \{Z_n\}_{n=1}^N \in \{1, 2, 3\}^N \)

**Graphical Model:**

- \( \{\pi_k\}_{k=1}^K \rightarrow Z_1 \)
- \( \{\mu_k, \Sigma_k\}_{k=1}^K \rightarrow X_1 \)
- \( Z_2 \rightarrow X_2 \)
- \( \ldots \)
- \( Z_N \rightarrow X_N \)

**Conditional:**
\[
p(X_n = x_n | Z_n = k; \theta) = \mathcal{N}(x_n; \mu_k, \Sigma_k)
\]

**Priors:**
\[
p(Z_n = k; \theta) = \pi_k, \quad \sum_{k=1}^K \pi_k = 1
\]
Bayesian Inference: Examples

Observed variables: \( \{x_n\}_{n=1}^{N} \subset \mathbb{R}^2 \)

Hidden Variables: \( \{Z_n\}_{n=1}^{N} \in \{1, 2, 3\}^N \)

Graphical Model:

- \( \{ \pi_k \}_{k=1}^{K} \)
- \( \{ \mu_k, \Sigma_k \}_{k=1}^{K} \)

Conditional:
\[
p(X_n = x_n | Z_n = k; \theta) = \mathcal{N}(x_n; \mu_k, \Sigma_k)
\]

Priors:
\[
p(Z_n = k; \theta) = \pi_k, \quad \sum_{k=1}^{K} \pi_k = 1
\]

Parameters:
\[
\theta = \{ \mu_k, \Sigma_k, \pi_k \}_{k=1}^{K}
\]
Bayesian Inference: Examples

Inference

Bayes’ Theorem: \( p(Z_1 = z_1, \ldots, Z_N = z_N | x_1, \ldots, x_N; \theta) = \prod_{n=1}^{N} p(Z_n = z_n | x_n; \theta) \)

where

\[
p(Z_n = k | x_n; \theta) = \frac{p(x_n | Z_n = k; \theta)p(Z_n = k; \theta)}{p(x_1, \ldots, x_N; \theta)} \propto \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)
\]
Inference

Bayes’ Theorem:

$$p(Z_1 = z_1, \ldots, Z_N = z_N | x_1, \ldots, x_N; \theta) = \prod_{n=1}^{N} p(Z_n = z_n | x_n; \theta)$$

where

$$p(Z_n = k | x_n; \theta) = \frac{p(x_n | Z_n = k; \theta)p(Z_n = k; \theta)}{p(x_1, \ldots, x_N; \theta)} \propto \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)$$

Simple form, but $\theta$ is unknown
Inference

Bayes’ Theorem: \( p(Z_1 = z_1, \ldots, Z_N = z_N | x_1, \ldots, x_N; \theta) = \prod_{n=1}^{N} p(Z_n = z_n | x_n; \theta) \)

where \( p(Z_n = k | x_n; \theta) = \frac{p(x_n | Z_n = k; \theta)p(Z_n = k; \theta)}{p(x_1, \ldots, x_N; \theta)} \propto \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k) \)

Simple form, but \( \theta \) is unknown \( \Rightarrow \) Maximum likelihood? \( \hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta; X) \)
Bayesian Inference: Examples

Inference

Bayes’ Theorem: \( p(Z_1 = z_1, \ldots, Z_N = z_N | x_1, \ldots, x_N; \theta) = \prod_{n=1}^{N} p(Z_n = z_n | x_n; \theta) \)

where \( p(Z_n = k | x_n; \theta) = \frac{p(x_n | Z_n = k; \theta)p(Z_n = k; \theta)}{p(x_1, \ldots, x_N; \theta)} \)

\( \propto \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k) \)

Simple form, but \( \theta \) is unknown \( \Rightarrow \) Maximum likelihood? \( \hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta; X) \)

\( \mathcal{L}(\theta; X) = p(X_1 = x_1, \ldots, X_N = x_N; \theta) \)

\( = \log \left( \sum_{z_1, \ldots, z_N = 1}^{K} \prod_{n=1}^{N} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k) \right) \)
Bayesian Inference: Examples

**Inference**

**Bayes’ Theorem:**
\[
p(Z_1 = z_1, \ldots, Z_N = z_N | x_1, \ldots, x_N; \theta) = \prod_{n=1}^{N} p(Z_n = z_n | x_n; \theta)
\]

where
\[
p(Z_n = k | x_n; \theta) = \frac{p(x_n | Z_n = k; \theta)p(Z_n = k; \theta)}{p(x_1, \ldots, x_N; \theta)}
\]

\[
\propto \pi_k N(x_n; \mu_k, \Sigma_k)
\]

Simple form, but \( \theta \) is unknown  \( \Rightarrow \) Maximum likelihood?
\[
\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta; X) = \arg\max_{\theta} \log \left( \sum_{z_1, \ldots, z_N=1}^{K} \prod_{n=1}^{N} \pi_k N(x_n; \mu_k, \Sigma_k) \right)
\]

\[
\mathcal{L}(\theta; X) = p(X_1 = x_1, \ldots, X_N = x_N; \theta)
\]

- Non-convex
- Combinatorial
- Intractable
**Bayesian Inference: Examples**

**Inference**

**Bayes’ Theorem:**

\[
p(Z_1 = z_1, \ldots, Z_N = z_N | x_1, \ldots, x_N; \theta) = \prod_{n=1}^{N} p(Z_n = z_n | x_n; \theta)
\]

where

\[
p(Z_n = k | x_n; \theta) = \frac{p(x_n | Z_n = k; \theta)p(Z_n = k; \theta)}{p(x_1, \ldots, x_N; \theta)} \propto \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)
\]

Simple form, but \( \theta \) is unknown  => Maximum likelihood?

\[
\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta; X)
\]

\[
\mathcal{L}(\theta; X) = p(X_1 = x_1, \ldots, X_N = x_N; \theta)
\]

\[
= \log \left( \sum_{z_1, \ldots, z_N=1}^{K} \prod_{n=1}^{N} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k) \right)
\]

- Non-convex
- Combinatorial
- Intractable

- The joint probability \( p(X, Z; \theta) \) has a much simpler form than the marginal \( p(X; \theta) \)

- \( Z \) is a hidden variable, and cannot be estimated without knowing \( \theta \)
Bayesian Inference: Examples

**Bayes’ Theorem:**
\[
p(Z_1 = z_1, \ldots, Z_N = z_N | x_1, \ldots, x_N; \theta) = \prod_{n=1}^{N} p(Z_n = z_n | x_n; \theta)
\]

where
\[
p(Z_n = k | x_n; \theta) = \frac{p(x_n | Z_n = k; \theta)p(Z_n = k; \theta)}{p(x_1, \ldots, x_N; \theta)}
\]
\[\propto \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k)
\]

Simple form, but \( \theta \) is unknown  \( \Rightarrow \) Maximum likelihood?
\[
\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta; X)
\]

\[
\mathcal{L}(\theta; X) = p(X_1 = x_1, \ldots, X_N = x_N; \theta)
\]
\[
= \log \left( \sum_{z_1, \ldots, z_N=1}^{K} \prod_{n=1}^{N} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k) \right)
\]

- Non-convex
- Combinatorial
- Intractable

\[
\{ \begin{align*}
\text{• The joint probability } & p(X, Z; \theta) \text{ has a much simpler form than the marginal } p(X; \theta) \\
\text{• } Z & \text{ is a hidden variable, and cannot be estimated without knowing } \theta
\end{align*}
\]

\[\Longrightarrow \text{ Expectation-Maximization (EM) algorithm}\]
EM algorithm

\[ Q(\theta|\theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}} \left[ \log \mathcal{L}(\theta; X, Z) \right] \]

• E-step: \( Q(\theta|\theta^{(i)}) \)

• M-step: \( \theta^{(i+1)} = \arg\max_{\theta} Q(\theta|\theta^{(i)}) \)

Complete-data log-likelihood
EM algorithm
\[
\begin{align*}
\text{• E-step: } & \quad Q(\theta|\theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}}[\log \mathcal{L}(\theta; X, Z)] \\
\text{• M-step: } & \quad \theta^{(i+1)} = \arg\max_{\theta} Q(\theta|\theta^{(i)})
\end{align*}
\]

Proof of correctness:
\[
p(Z|X, \theta) = \frac{p(X, Z|\theta)}{p(X|\theta)} \quad \text{(Baye’s theorem)}
\]
Bayesian Inference: Examples

EM algorithm

\[ \begin{align*}
\text{• E-step: } Q(\theta|\theta^{(i)}) &= \mathbb{E}_{Z|X,\theta^{(i)}}[\log \mathcal{L}(\theta; X, Z)] \\
\text{• M-step: } \theta^{(i+1)} &= \arg\max_{\theta} Q(\theta|\theta^{(i)})
\end{align*} \]

Proof of correctness:

\[ p(Z|X, \theta) = \frac{p(X, Z|\theta)}{p(X|\theta)} \] (Baye’s theorem)

\[ \log p(X|\theta) = \log p(X, Z|\theta) - \log p(Z|X, \theta) \]
EM algorithm

\[ \begin{align*}
\text{E-step:} & \quad Q(\theta|\theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}} \left[ \log \mathcal{L}(\theta; X, Z) \right] \\
\text{M-step:} & \quad \theta^{(i+1)} = \arg\max_{\theta} Q(\theta|\theta^{(i)})
\end{align*} \]

Proof of correctness:

\[ p(Z|X, \theta) = \frac{p(X, Z|\theta)}{p(X|\theta)} \quad \text{(Baye's theorem)} \]

\[ \log p(X|\theta) = \log p(X, Z|\theta) - \log p(Z|X, \theta) \]

\[ \mathbb{E}_{Z|X, \theta^{(i)}} [\log p(X|\theta)] = \sum_{Z} p(Z|X, \theta^{(i)}) \log p(X, Z|\theta) - \sum_{Z} p(Z|X, \theta^{(i)}) \log p(Z|X, \theta) \]
EM algorithm

\[ Q(\theta|\theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}}[\log \mathcal{L}(\theta; X, Z)] \]

\[ \theta^{(i+1)} = \arg\max_{\theta} Q(\theta|\theta^{(i)}) \]

Proof of correctness:

\[ p(Z|X, \theta) = \frac{p(X, Z|\theta)}{p(X|\theta)} \quad \text{(Baye’s theorem)} \]

\[ \log p(X|\theta) = \log p(X, Z|\theta) - \log p(Z|X, \theta) \]

\[ \mathbb{E}_{Z|X, \theta^{(i)}}[\log p(X|\theta)] = \sum_Z p(Z|X, \theta^{(i)}) \log p(X, Z|\theta) - \sum_Z p(Z|X, \theta^{(i)}) \log p(Z|X, \theta) \]

\[ \log p(X|\theta) = \mathcal{L}(\theta; X) = Q(\theta|\theta^{(i)}) + H(\theta|\theta^{(i)}) \]

where \( H(\theta|\theta^{(i)}) \) is the conditional cross entropy of \( Z \) given \( X, \theta \) for the distribution \( p(Z|X, \theta^{(i)}) \).
Bayesian Inference: Examples

EM algorithm

\[ Q(\theta | \theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}}[\log \mathcal{L}(\theta; X, Z)] \]

- **E-step:**

- **M-step:** \( \theta^{(i+1)} = \arg\max_{\theta} Q(\theta | \theta^{(i)}) \)

Proof of correctness:

\[
p(Z|X, \theta) = \frac{p(X, Z|\theta)}{p(X|\theta)} \quad \text{(Baye’s theorem)}
\]

\[
\log p(X|\theta) = \log p(X, Z|\theta) - \log p(Z|X, \theta)
\]

\[
\mathbb{E}_{Z|X,\theta^{(i)}}[\log p(X|\theta)] = \sum_{Z} p(Z|X, \theta^{(i)}) \log p(X, Z|\theta) - \sum_{Z} p(Z|X, \theta^{(i)}) \log p(Z|X, \theta)
\]

\[
\log p(X|\theta) = \mathcal{L}(\theta; X) = Q(\theta|\theta^{(i)}) + H(\theta|\theta^{(i)})
\]

where \( H(\theta|\theta^{(i)}) \) is the conditional cross entropy of \( Z \) given \( X, \theta \) for the distribution \( p(Z|X, \theta^{(i)}) \).

In particular:

\[
\mathcal{L}(\theta^{(i)}; X) = Q(\theta^{(i)}|\theta^{(i)}) + H(\theta^{(i)}|\theta^{(i)}).
\]
Bayesian Inference: Examples

EM algorithm

\( \begin{cases} \text{• E-step: } Q(\theta | \theta^{(i)}) = \mathbb{E}_{Z | X, \theta^{(i)}} [\log \mathcal{L}(\theta; X, Z)] \\ \text{• M-step: } \theta^{(i+1)} = \arg \max_{\theta} Q(\theta | \theta^{(i)}) \end{cases} \)

Proof of correctness:

\[
p(Z | X, \theta) = \frac{p(X, Z | \theta)}{p(X | \theta)} \quad \text{(Baye’s theorem)}
\]

\[
\log p(X | \theta) = \log p(X, Z | \theta) - \log p(Z | X, \theta)
\]

\[
\mathbb{E}_{Z | X, \theta^{(i)}} [\log p(X | \theta)] = \sum_Z p(Z | X, \theta^{(i)}) \log p(X, Z | \theta) - \sum_Z p(Z | X, \theta^{(i)}) \log p(Z | X, \theta)
\]

\[
\log p(X | \theta) = \mathcal{L}(\theta; X) = Q(\theta | \theta^{(i)}) + H(\theta | \theta^{(i)})
\]

where \( H(\theta | \theta^{(i)}) \) is the conditional cross entropy of \( Z \) given \( X, \theta \) for the distribution \( p(Z | X, \theta^{(i)}) \).

In particular:

\[
\mathcal{L}(\theta^{(i)}; X) = Q(\theta^{(i)} | \theta^{(i)}) + H(\theta^{(i)} | \theta^{(i)}).
\]

Therefore:

\[
\mathcal{L}(\theta; X) - \mathcal{L}(\theta^{(i)}; X) = Q(\theta | \theta^{(i)}) - Q(\theta^{(i)} | \theta^{(i)}) + H(\theta | \theta^{(i)}) - H(\theta^{(i)} | \theta^{(i)}).
\]
EM algorithm

\[ Q(\theta | \theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}} [\log L(\theta; X, Z)] \]

\[ \theta^{(i+1)} = \operatorname{argmax}_\theta Q(\theta | \theta^{(i)}) \]

Proof of correctness:

\[ p(Z|X, \theta) = \frac{p(X, Z|\theta)}{p(X|\theta)} \] (Baye’s theorem)

\[ \log p(X|\theta) = \log p(X, Z|\theta) - \log p(Z|X, \theta) \]

\[ \mathbb{E}_{Z|X, \theta^{(i)}} [\log p(X|\theta)] = \sum_Z p(Z|X, \theta^{(i)}) \log p(X, Z|\theta) - \sum_Z p(Z|X, \theta^{(i)}) \log p(Z|X, \theta) \]

\[ \log p(X|\theta) = \mathcal{L}(\theta; X) = Q(\theta|\theta^{(i)}) + H(\theta|\theta^{(i)}) \]

where \( H(\theta|\theta^{(i)}) \) is the conditional cross entropy of \( Z \) given \( X, \theta \) for the distribution \( p(Z|X, \theta^{(i)}) \).

In particular:

\[ \mathcal{L}(\theta^{(i)}; X) = Q(\theta^{(i)}|\theta^{(i)}) + H(\theta^{(i)}|\theta^{(i)}) \]

Therefore:

\[ \mathcal{L}(\theta; X) - \mathcal{L}(\theta^{(i)}; X) = Q(\theta|\theta^{(i)}) - Q(\theta^{(i)}|\theta^{(i)}) + H(\theta|\theta^{(i)}) - H(\theta^{(i)}|\theta^{(i)}) \]

According to Gibb’s inequality, we have \( H(\theta|\theta^{(i)}) \geq H(\theta^{(i)}|\theta^{(i)}) \).
EM algorithm

\[ Q(\theta | \theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}} [\log \mathcal{L}(\theta; X, Z)] \]

\[ \theta^{(i+1)} = \arg\max_{\theta} Q(\theta | \theta^{(i)}) \]

Proof of correctness:

\[ p(Z|X, \theta) = \frac{p(X, Z|\theta)}{p(X|\theta)} \quad \text{(Baye’s theorem)} \]

\[ \log p(X|\theta) = \log p(X, Z|\theta) - \log p(Z|X, \theta) \]

\[ \mathbb{E}_{Z|X, \theta^{(i)}} [\log p(X|\theta)] = \sum_{Z} p(Z|X, \theta^{(i)}) \log p(X, Z|\theta) - \sum_{Z} p(Z|X, \theta^{(i)}) \log p(Z|X, \theta) \]

\[ \log p(X|\theta) = \mathcal{L}(\theta; X) = Q(\theta | \theta^{(i)}) + H(\theta | \theta^{(i)}) \]

where \( H(\theta | \theta^{(i)}) \) is the conditional cross entropy of \( Z \) given \( X, \theta \) for the distribution \( p(Z|X, \theta^{(i)}) \).

In particular:

\[ \mathcal{L}(\theta^{(i)}; X) = Q(\theta^{(i)} | \theta^{(i)}) + H(\theta^{(i)} | \theta^{(i)}) \]

Therefore:

\[ \mathcal{L}(\theta; X) - \mathcal{L}(\theta^{(i)}; X) = Q(\theta | \theta^{(i)}) - Q(\theta^{(i)} | \theta^{(i)}) + H(\theta | \theta^{(i)}) - H(\theta^{(i)} | \theta^{(i)}) \]

According to Gibb’s inequality, we have \( H(\theta | \theta^{(i)}) \geq H(\theta^{(i)} | \theta^{(i)}) \).

\[ \mathcal{L}(\theta; X) - \mathcal{L}(\theta^{(i)}; X) \geq Q(\theta | \theta^{(i)}) - Q(\theta^{(i)} | \theta^{(i)}) \]
EM algorithm

\[ \begin{align*}
\text{E-step: } & \quad Q(\theta | \theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}}[\log \mathcal{L}(\theta; X, Z)] \\
\text{M-step: } & \quad \theta^{(i+1)} = \arg \max_{\theta} Q(\theta | \theta^{(i)})
\end{align*} \]

Proof of correctness:

\[
\begin{align*}
p(Z|X, \theta) &= \frac{p(X, Z|\theta)}{p(X|\theta)} \quad \text{(Baye’s theorem)} \\
\log p(X|\theta) &= \log p(X, Z|\theta) - \log p(Z|X, \theta) \\
\mathbb{E}_{Z|X, \theta^{(i)}}[\log p(X|\theta)] &= \sum_Z p(Z|X, \theta^{(i)}) \log p(X, Z|\theta) - \sum_Z p(Z|X, \theta^{(i)}) \log p(Z|X, \theta) \\
\log p(X|\theta) &= \mathcal{L}(\theta; X) = Q(\theta|\theta^{(i)}) + H(\theta|\theta^{(i)})
\end{align*}\]

where \( H(\theta|\theta^{(i)}) \) is the conditional cross entropy of \( Z \) given \( X, \theta \) for the distribution \( p(Z|X, \theta^{(i)}) \).

In particular:

\[ \mathcal{L}(\theta^{(i)}; X) = Q(\theta^{(i)}|\theta^{(i)}) + H(\theta^{(i)}|\theta^{(i)}) \]

Therefore:

\[ \mathcal{L}(\theta; X) - \mathcal{L}(\theta^{(i)}; X) = Q(\theta|\theta^{(i)}) - Q(\theta^{(i)}|\theta^{(i)}) + H(\theta|\theta^{(i)}) - H(\theta^{(i)}|\theta^{(i)}) \]

According to Gibb’s inequality, we have \( H(\theta|\theta^{(i)}) \geq H(\theta^{(i)}|\theta^{(i)}) \).

The likelihood can only increase at each step!
Derivations for the Gaussian mixture model
Derivations for the Gaussian mixture model

• **E-step:** computing the current posterior probabilities

\[
    r_{n,k}^{(i)} = p(Z_n = k | x_n; \theta^{(i)}) = \frac{p(x_n | Z_n = k; \theta^{(i)}) p(Z_n = k; \theta^{(i)})}{p(x_1, \ldots, x_N; \theta^{(i)})} \propto \pi_k^{(i)} \mathcal{N}(x_n; \mu_k^{(i)}, \Sigma_k^{(i)})
\]
Derivations for the Gaussian mixture model

**E-step:** computing the current posterior probabilities

\[ r_{n,k}^{(i)} = p(Z_n = k | x_n; \theta^{(i)}) = \frac{p(x_n | Z_n = k; \theta^{(i)}) p(Z_n = k; \theta^{(i)})}{p(x_1, \ldots, x_N; \theta^{(i)})} \propto \pi_k^{(i)} \mathcal{N}(x_n; \mu_k^{(i)}, \Sigma_k^{(i)}) \]

We deduce \( Q(\theta | \theta^{(i)}) \):

\[
Q(\theta | \theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}} \{ \log \mathcal{L}(\theta; X, Z) \}
\]

\[
= \mathbb{E}_{Z|X,\theta^{(i)}} \left\{ \log \prod_{n=1}^{N} \sum_{k=1}^{K} \mathbb{I}\{Z_n = k\} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k) \right\}
\]

\[
= \sum_{k,n=1}^{K,N} \mathbb{E}_{Z|X,\theta^{(i)}} \{ \mathbb{I}\{Z_n = k\} \} \log (\pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k))
\]

\[
= \sum_{k,n=1}^{K,N} r_{n,k}^{(i)} \left( \log \pi_k - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_n - \mu_k)^\top \Sigma_k^{-1} (x_n - \mu_k) \right) + \text{const.}
\]
Derivations for the Gaussian mixture model

• **E-step:** computing the current posterior probabilities

\[
r_{n,k}^{(i)} = p(Z_n = k | x_n; \theta^{(i)}) = \frac{p(x_n | Z_n = k; \theta^{(i)}) p(Z_n = k; \theta^{(i)})}{p(x_1, \ldots, x_N; \theta^{(i)})} \propto \pi_k^{(i)} \mathcal{N}(x_n; \mu_k^{(i)}, \Sigma_k^{(i)})
\]

We deduce \( Q(\theta | \theta^{(i)}) \):\[
Q(\theta | \theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}} \{ \log \mathcal{L}(\theta; X, Z) \}
\]
\[
= \mathbb{E}_{Z|X,\theta^{(i)}} \left\{ \log \prod_{n=1}^{N} \sum_{k=1}^{K} \mathbb{I}\{Z_n = k\} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k) \right\}
\]
\[
= \sum_{k,n=1}^{K,N} \mathbb{E}_{Z|X,\theta^{(i)}} \{ \mathbb{I}\{Z_n = k\} \} \log (\pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k))
\]
\[
= \sum_{k,n=1}^{K,N} r_{n,k}^{(i)} \left( \log \pi_k - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_n - \mu_k)^\top \Sigma_k^{-1}(x_n - \mu_k) \right) + \text{const.}
\]

• **M-step:** maximizing \( Q(\theta | \theta^{(i)}) \) by finding the zeros of the derivative
Bayesian Inference: Examples

Derivations for the Gaussian mixture model

• **E-step:** computing the current posterior probabilities

\[
r^{(i)}_{n,k} = p(Z_n = k | x_n; \theta^{(i)}) = \frac{p(x_n | Z_n = k; \theta^{(i)}) p(Z_n = k; \theta^{(i)})}{p(x_1, \ldots, x_N; \theta^{(i)})} \propto \pi^{(i)}_k \mathcal{N}(x_n; \mu^{(i)}_k, \Sigma^{(i)}_k)
\]

We deduce \( Q(\theta | \theta^{(i)}) \):

\[
Q(\theta | \theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}} \{ \log \mathcal{L}(\theta; X, Z) \}
\]

\[
= \mathbb{E}_{Z|X,\theta^{(i)}} \left\{ \log \prod_{n=1}^{N} \sum_{k=1}^{K} \mathbb{I}\{Z_n = k\} \pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k) \right\}
\]

\[
= \sum_{k, n=1}^{K, N} \mathbb{E}_{Z|X,\theta^{(i)}} \{ \mathbb{I}\{Z_n = k\} \} \log (\pi_k \mathcal{N}(x_n; \mu_k, \Sigma_k))
\]

\[
= \sum_{k, n=1}^{K, N} r^{(i)}_{n,k} (\log \pi_k - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x_n - \mu_k)^\top \Sigma_k^{-1} (x_n - \mu_k)) + \text{const.}
\]

• **M-step:** maximizing \( Q(\theta | \theta^{(i)}) \) by finding the zeros of the derivative

\[
\pi^{(i+1)}_k = \frac{1}{N} \sum_{n=1}^{N} r^{(i)}_{n,k}, \quad \mu^{(i+1)}_k = \frac{1}{N} \sum_{n=1}^{N} r^{(i)}_{n,k} x_n, \quad \Sigma^{(i+1)}_k = \frac{1}{N} \sum_{n=1}^{N} r^{(i)}_{n,k} (x_n - \mu^{(i+1)}_k) (x_n - \mu^{(i+1)}_k)^\top
\]
EM algorithm

- Initialization: Random "guess" for $\theta$
- **E-step:** $Q(\theta | \theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}}[\log \mathcal{L}(\theta; X, Z)]$
- **M-step:** $\theta^{(i+1)} = \arg\max_\theta Q(\theta | \theta^{(i)})$
- Convergence
Bayesian Inference: Examples

EM algorithm

\begin{itemize}
  \item **Initialization:** Random « guess » for $\theta$
  \item **E-step:** $Q(\theta | \theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}} \{ \log \mathcal{L}(\theta; X, Z) \}$
  \item **M-step:** $\theta^{(i+1)} = \arg\max_{\theta} Q(\theta | \theta^{(i)})$
  \item **Convergence**
\end{itemize}
Bayesian Inference: Examples

**Inference**

**EM algorithm**

- **Initialization**: Random « guess » for $\theta$

- **E-step**: 
  $$Q(\theta | \theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}}[\log \mathcal{L}(\theta; X, Z)]$$

- **M-step**: 
  $$\theta^{(i+1)} = \arg \max_{\theta} Q(\theta | \theta^{(i)})$$

- **Convergence**
Bayesian Inference: Examples

EM algorithm

Inference

\[ Q(\theta|\theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}}[\log \mathcal{L}(\theta; X, Z)] \]

\[ \theta^{(i+1)} = \arg\max_{\theta} Q(\theta|\theta^{(i)}) \]

- Initialization: Random « guess » for \( \theta \)
- E-step: \( Q(\theta|\theta^{(i)}) = \) ...
- M-step: \( \theta^{(i+1)} = \) ...
- Convergence
Bayesian Inference: Examples

EM algorithm

Inference

\begin{itemize}
\item Initialization: Random « guess » for $\theta$
\item E-step: $Q(\theta|\theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}}[\log \mathcal{L}(\theta; X, Z)]$
\item M-step: $\theta^{(i+1)} = \arg\max_{\theta} Q(\theta|\theta^{(i)})$
\item Convergence
\end{itemize}
Bayesian Inference: Examples

Inference

**EM algorithm**

- **Initialization**: Random « guess » for $\theta$
- **E-step**: $Q(\theta|\theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}} \left[ \log \mathcal{L}(\theta; X, Z) \right]$
- **M-step**: $\theta^{(i+1)} = \arg\max_{\theta} Q(\theta|\theta^{(i)})$
- **Convergence**
Bayesian Inference: Examples

EM algorithm

- **Inference**
  - **EM algorithm**
    - **Initialization**: Random « guess » for $\theta$
    - **E-step**: $Q(\theta|\theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}}[\log \mathcal{L}(\theta; X, Z)]$
    - **M-step**: $\theta^{(i+1)} = \arg \max_{\theta} Q(\theta|\theta^{(i)})$
    - **Convergence**
EM algorithm

- **Initialization**: Random «guess» for $\theta$
- **E-step**: $Q(\theta|\theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}}[\log \mathcal{L}(\theta; X, Z)]$
- **M-step**: $\theta^{(i+1)} = \operatorname{argmax}_\theta Q(\theta|\theta^{(i)})$
- **Convergence**
EM algorithm

• Initialization: Random « guess » for $\theta$

• E-step: $Q(\theta|\theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}}[\log \mathcal{L}(\theta; X, Z)]$

• M-step: $\theta^{(i+1)} = \arg\max_{\theta} Q(\theta|\theta^{(i)})$

• Convergence
Bayesian Inference: Examples

Inference

EM algorithm

- Initialization: Random « guess » for $\theta$
- E-step: $Q(\theta|\theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}}[\log L(\theta; X, Z)]$
- M-step: $\theta^{(i+1)} = \arg\max_\theta Q(\theta|\theta^{(i)})$
- Convergence

EM algorithm
Bayesian Inference: Examples

Inference

EM algorithm

\[
\begin{align*}
\text{• Initialization: Random "guess" for } \theta \\
\text{• E-step: } Q(\theta|\theta^{(i)}) &= \mathbb{E}_{Z|X,\theta^{(i)}}[\log \mathcal{L}(\theta; X, Z)] \\
\text{• M-step: } \theta^{(i+1)} &= \arg\max_{\theta} Q(\theta|\theta^{(i)}) \\
\text{• Convergence}
\end{align*}
\]
**Inference**

**EM algorithm**

- **Initialization**: Random « guess » for $\theta$
- **E-step**: $Q(\theta|\theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}}[\log \mathcal{L}(\theta; X, Z)]$
- **M-step**: $\theta^{(i+1)} = \arg\max_{\theta} Q(\theta|\theta^{(i)})$

**Convergence**
EM algorithm

- **Initialization**: Random « guess » for $\theta$
- **E-step**: $Q(\theta|\theta^{(i)}) = \mathbb{E}_{Z|X,\theta^{(i)}}[\log \mathcal{L}(\theta; X, Z)]$
- **M-step**: $\theta^{(i+1)} = \arg\max_{\theta} Q(\theta|\theta^{(i)})$

- **Convergence**

Decision:

Minus 10 points for Mr. Green, minus 5 points for the others!
• What is Bayesian inference?
  – Overview
  – Classical vs. Bayesian approach
  – Bayes Theorem & Example
  – General Methodology

• Bayesian inference by examples
  – Direct inference
  – The Expectation-Maximization algorithm
  – Variational Bayes methods

• Markov chain Monte Carlo
• What is Bayesian inference?
  – Overview
  – Classical vs. Bayesian approach
  – Bayes Theorem & Example
  – General Methodology

• Bayesian inference by examples
  – Direct inference
  – The Expectation-Maximization algorithm
  – Variational Bayes methods

• Markov chain Monte Carlo
The next day...
The next day...
The next day...
The next day...
The next day...
The next day...
Bayesian Inference: Examples
I will show them what Bayes is capable of...
A « fully Bayesian » model

\[
p(x_n|Z_n = k, \mu_k, \Lambda_k) = \mathcal{N}(x_n; \mu_k, \Lambda_k^{-1})
\]
\[
p(Z_n = k|\pi) = \pi_k
\]

GMM (same as before)
A « fully Bayesian » model

\[
p(x_n|Z_n = k, \mu_k, \Lambda_k) = \mathcal{N}(x_n; \mu_k, \Lambda_k^{-1})
\]

\[
p(Z_n = k|\pi) = \pi_k
\]

\[
p(\Lambda_k) = \mathcal{W}(\Lambda_k; W_0, \nu_0)
\]

\[
p(\mu_k|\Lambda_k) = \mathcal{N}(\mu_k; m_0, \Lambda_k^{-1})
\]

\[
p(\pi) = \text{SymDir}(\pi|\alpha_0)
\]

GMM (same as before)

Priors on all parameters

Note: These are the conjugate priors for the normal and the multinomial distributions, i.e., they are such that

\[
p(\mu_k, \Lambda_k|x_n, Z_n = k) \cong p(\mu_k, \Lambda_k) \text{ and } p(\pi|z_n) \cong p(\pi)
\]
Bayesian Inference: Examples

Modeling

A « fully Bayesian » model

\[
p(x_n|Z_n = k, \mu_k, \Lambda_k) = \mathcal{N}(x_n; \mu_k, \Lambda_k^{-1}) \]
\[
p(Z_n = k|\pi) = \pi_k \]

GMM (same as before)

\[
p(\Lambda_k) = \mathcal{W}(\Lambda_k; \mathbf{W}_0, \nu_0) \]
\[
p(\mu_k|\Lambda_k) = \mathcal{N}(\mu_k; \mathbf{m}_0, \Lambda_k^{-1}) \]
\[
p(\pi) = \text{SymDir}(\pi|\alpha_0) \]

Priors on all parameters

Note: These are the conjugate priors for the normal and the multinomial distributions, i.e., they are such that

\[
p(\mu_k, \Lambda_k|x_n, Z_n = k) \approx p(\mu_k, \Lambda_k) \text{ and } p(\pi|z_n) \approx p(\pi)\]
Bayesian Inference: Examples

Modeling

A « fully Bayesian » model

\[
p(x_n|Z_n = k, \mu_k, \Lambda_k) = \mathcal{N}(x_n; \mu_k, \Lambda_k^{-1}) \\
p(Z_n = k|\pi) = \pi_k
\]

\[
p(\Lambda_k) = \mathcal{W}(\Lambda_k; W_0, \nu_0) \\
p(\mu_k|\Lambda_k) = \mathcal{N}(\mu_k; m_0, \Lambda_k^{-1}) \\
p(\pi) = \text{SymDir}(\pi|\alpha_0)
\]

GMM (same as before)

Priors on all parameters

Note: These are the conjugate priors for the normal and the multinomial distributions, i.e., they are such that

\[
p(\mu_k, \Lambda_k|x_n, Z_n = k) \approx p(\mu_k, \Lambda_k) \quad \text{and} \quad p(\pi|z_n) \approx p(\pi)
\]

Graphical model:

Choice of hyperparameters:

\[
W_0 = I \\
\nu_0 = D = 2 \\
m_0 = \text{mean}(X) \\
\alpha_0 > 0: \text{low values will allow Gaussian weights to be close to 0}
\]
Inference

- The posterior distribution $p(Z, \Lambda, \mu, \pi | X)$ is intractable
Inference

• The posterior distribution \( p(Z, \Lambda, \mu, \pi | X) \) is intractable

• **Technique**: use a *variational approximation* \( p(Z, \Lambda, \mu, \pi | X) \approx q(Z, \Lambda, \mu, \pi) \), where \( q(Z, \Lambda, \mu, \pi) \) is restricted to a family of distributions having a simpler form than the true \( p(Z, \Lambda, \mu, \pi | X) \)
Inference

• The posterior distribution $p(Z, \Lambda, \mu, \pi | X)$ is intractable

• **Technique**: use a variational approximation $p(Z, \Lambda, \mu, \pi | X) \approx q(Z, \Lambda, \mu, \pi)$, where $q(Z, \Lambda, \mu, \pi)$ is restricted to a family of distributions having a simpler form than the true $p(Z, \Lambda, \mu, \pi | X)$

• The variational distribution $q(Z, \Lambda, \mu, \pi)$ is typically assumed to **factorize** over some partition of the latent variables.
Inference

- The posterior distribution $p(Z, \Lambda, \mu, \pi | X)$ is intractable

- **Technique**: use a *variational approximation* $p(Z, \Lambda, \mu, \pi | X) \approx q(Z, \Lambda, \mu, \pi)$, where $q(Z, \Lambda, \mu, \pi)$ is restricted to a family of distributions having a simpler form than the true $p(Z, \Lambda, \mu, \pi | X)$

- The variational distribution $q(Z, \Lambda, \mu, \pi)$ is typically assumed to **factorize** over some partition of the latent variables.

- Here we use: $q(Z, \Lambda, \mu, \pi) = q_Z(Z)q_{\Lambda,\mu,\pi}(\Lambda, \mu, \pi)$. Remarkably, this is the **only** assumption needed to obtain a tractable *EM-like* inference procedure.
Inference

• The posterior distribution \( p(Z, \Lambda, \mu, \pi | X) \) is intractable

• **Technique:** use a *variational approximation* \( p(Z, \Lambda, \mu, \pi | X) \approx q(Z, \Lambda, \mu, \pi) \), where \( q(Z, \Lambda, \mu, \pi) \) is restricted to a family of distributions having a simpler form than the true \( p(Z, \Lambda, \mu, \pi | X) \)

• The variational distribution \( q(Z, \Lambda, \mu, \pi) \) is typically assumed to *factorize* over some partition of the latent variables.

• Here we use: \( q(Z, \Lambda, \mu, \pi) = q_Z(Z)q_{\Lambda,\mu,\pi}(\Lambda, \mu, \pi) \). Remarkably, this is the only assumption needed to obtain a tractable *EM-like* inference procedure.

• Such procedures are referred to as *Variational Bayesian EM algorithms.*
Bayesian Inference: Examples

Inference

- The posterior distribution $p(Z, \Lambda, \mu, \pi | X)$ is intractable

- **Technique**: use a *variational approximation* $p(Z, \Lambda, \mu, \pi | X) \approx q(Z, \Lambda, \mu, \pi)$, where $q(Z, \Lambda, \mu, \pi)$ is restricted to a family of distributions having a simpler form than the true $p(Z, \Lambda, \mu, \pi | X)$

- The variational distribution $q(Z, \Lambda, \mu, \pi)$ is typically assumed to **factorize** over some partition of the latent variables.

- Here we use: $q(Z, \Lambda, \mu, \pi) = q_Z(Z)q_{\Lambda\mu\pi}(\Lambda, \mu, \pi)$. Remarkably, this is the only assumption needed to obtain a tractable *EM-like* inference procedure.

- Such procedures are referred to as **Variational Bayesian EM algorithms**.
  For $p(Z, W | X) \approx q_Z(Z)q_W(W)$:

  \[
  \begin{align*}
  \text{VB-EM} & \quad \begin{cases} 
    \text{E-Z step:} & q_Z^{(i)}(Z) \propto \exp \left( \mathbb{E}_{q_W^{(i-1)}(W)} \{ \log p(Z | X, W) \} \right) \\
    \text{E-W step:} & q_W^{(i)}(W) \propto \exp \left( \mathbb{E}_{q_Z^{(i)}(Z)} \{ \log p(W | X, Z) \} \right)
  \end{cases}
  \end{align*}
  \]
Inference

Proof of correctness
• Using a similar reasoning as for EM, we can show that the VB-EM iteratively minimizes the Kulback-Leibler divergence between the true posterior $p(Z, W | X)$ and its variational approximation $q(Z, W) = q_Z(Z)q_W(W)$:

$$(q_Z^{(\infty)}, q_W^{(\infty)}) = \arg\min_{q_Z, q_W} KL(q || p)$$

VB-EM

\[\begin{align*}
\text{E-Z step: } & q_Z^{(i)}(Z) \propto \exp\left(\mathbb{E}_{q_W^{(i-1)}(W)}\{\log p(Z | X, W)\}\right) \\
\text{E-W step: } & q_W^{(i)}(W) \propto \exp\left(\mathbb{E}_{q_Z^{(i)}(Z)}\{\log p(W | X, Z)\}\right)
\end{align*}\]
Derivations for the Bayesian mixture of Gaussian

**E-Λμπ-Step:**

\[
\log q^{(i)}_{Λ, μ, π}(Λ, μ, π) = \log p(π) + \sum_{k=1}^{K} \log p(μ_k, Λ_k) + \sum_{n,k=1}^{N,K} q^{(i-1)}_{Z_n}(k) \log π_k N(x_n; μ_k, Λ_k^{-1}) + \text{const.}
\]

Using the decomposition \( q^{(i-1)}_{Z}(Z) = \prod_{n=1}^{N} q^{(i-1)}_{Z_n}(k) \) (see E-Z-step).
Derivations for the Bayesian mixture of Gaussian

**E-Λμπ-Step:**

\[
\log q_{Λ,μ,π}^{(i)}(Λ, μ, π) = \log p(π) + \sum_{k=1}^{K} \log p(μ_k, Λ_k) + \sum_{n,k=1}^{N,K} q_{Z_n}^{(i-1)}(k) \log π_k N(x_n; μ_k, Λ_k^{-1}) + \text{const.}
\]

Using the decomposition \( q_{Z}^{(i-1)}(Z) = \prod_{n=1}^{N} q_{Z_n}^{(i-1)}(k) \) (see E-Z-step).

This leads to the factorization:

\[
q_{Λ,μ,π}^{(i)}(Λ, μ, π) = q_π^{(i)}(π) \prod_{k=1}^{K} q_{Λ_kμ_k}^{(i)}(Λ_k, μ_k)
\]
Bayesian Inference: Examples

Inference

Derivations for the Bayesian mixture of Gaussian

• E-Λμπ-Step:

\[
\log q^{(i)}_{\Lambda, \mu, \pi}(\Lambda, \mu, \pi) = \log p(\pi) + \sum_{k=1}^{K} \log p(\mu_k, \Lambda_k) + \sum_{n,k=1}^{N,K} q_{Z_n}^{(i-1)}(k) \log \pi_k \mathcal{N}(x_n; \mu_k, \Lambda_k^{-1}) + \text{const.}
\]

Using the decomposition \( q_{Z_n}^{(i-1)}(k) = \prod_{n=1}^{N} q_{Z_n}^{(i-1)}(k) \) (see E-Z-step).

This leads to the factorization:

\[
q^{(i)}_{\Lambda, \mu, \pi}(\Lambda, \mu, \pi) = q^{(i)}_{\pi}(\pi) \prod_{k=1}^{K} q^{(i)}_{\Lambda_k \mu_k}(\Lambda_k, \mu_k)
\]

with

\[
\begin{align*}
q^{(i)}_{\pi}(\pi) &= \text{Dir}(\pi; \alpha_0 + N^{(i)}_1, \ldots, \alpha_0 + N^{(i)}_K) \\
q^{(i)}_{\Lambda_k \mu_k}(\Lambda_k, \mu_k) &= \mathcal{N}
\left(
\mu_k; m^{(i)}_k, \frac{\Lambda_k^{-1}}{1 + N^{(i)}_k}
\right)
\mathcal{W}
\left(
\Lambda_k; W^{(i)}_k, \nu^{(i)}_k
\right)
\end{align*}
\]

Variational methods
Derivations for the Bayesian mixture of Gaussian

**E-Λμπ-Step:**

\[
\log q_{\Lambda \mu \pi}^{(i)}(\Lambda, \mu, \pi) = \log p(\pi) + \sum_{k=1}^{K} \log p(\mu_k, \Lambda_k) + \sum_{n,k=1}^{N,K} q_{Z_n}^{(i-1)}(k) \log \pi_k \mathcal{N}(x_n; \mu_k, \Lambda_k^{-1}) + \text{const}.
\]

Using the decomposition \( q_{Z_n}^{(i-1)}(Z) = \prod_{n=1}^{N} q_{Z_n}^{(i-1)}(k) \) (see E-Z-step).

This leads to the factorization:

\[
q_{\Lambda \mu \pi}^{(i)}(\Lambda, \mu, \pi) = q_{\pi}^{(i)}(\pi) \prod_{k=1}^{K} q_{\Lambda_k \mu_k}^{(i)}(\Lambda_k, \mu_k)
\]

with

\[
\begin{align*}
q_{\pi}^{(i)}(\pi) &= \text{Dir}(\pi; \alpha_0 + N_1^{(i)}, \ldots, \alpha_0 + N_K^{(i)}) \\
q_{\Lambda_k \mu_k}^{(i)}(\Lambda_k, \mu_k) &= \mathcal{N}\left(\mu_k; m_k^{(i)}, \frac{\Lambda_k^{-1}}{1 + N_k^{(i)}}\right) \mathcal{W}\left(\Lambda_k; W_k^{(i)}, \nu_k^{(i)}\right),
\end{align*}
\]

where:

\[
\bar{x}_k^{(i)} = \frac{1}{N_k^{(i)}} \sum_{n=1}^{N} q_{Z_n}^{(i-1)}(k) x_n,
\]

\[
S_k^{(i)} = \frac{1}{N_k^{(i)}} \sum_{n=1}^{N} q_{Z_n}^{(i-1)}(k) (x_n - \bar{x}_k^{(i)}) (x_n - \bar{x}_k^{(i)})^T,
\]

\[
N_k^{(i)} = \sum_{n=1}^{N} q_{Z_n}^{(i-1)}(k),
\]

\[
m_k^{(i)} = \frac{m_0 + N_k^{(i)} \bar{x}_k^{(i)}}{1 + N_k^{(i)}},
\]

\[
W_k^{(i)-1} = W_0^{-1} + N_k^{(i)} S_k^{(i)} + \frac{N_k^{(i)}}{N_k^{(i)} + 1} (\bar{x}_k^{(i)} - m_0) (\bar{x}_k^{(i)} - m_0)^T,
\]

\[
\nu_k^{(i)} = \nu_0 + N_k^{(i)}
\]
Derivations for the Bayesian mixture of Gaussian

• E-Z-Step:
Derivations for the Bayesian mixture of Gaussian

• **E-Z-Step:**

\[
\log q_Z^{(i)}(\mathbf{Z}) = \mathbb{E}_{q_{\Lambda^\mu\pi}}^{(i)} \{ \log p(\mathbf{Z} | \mathbf{X}, \Lambda, \mu, \pi) \} + \text{const.}
\]

\[
= \sum_{n,k=1}^{N,K} \mathbb{I}\{Z_n = k\} \log \rho_{n,k}^{(i)} + \text{const.}
\]

where

\[
\rho_{n,k}^{(i)} = \mathbb{E}_{q_{\pi}}^{(i)} \{ \log \pi_k \} + \frac{1}{2} \mathbb{E}_{q_{\Lambda_k}}^{(i)} \{ \log |\Lambda_k| \} - \frac{D}{2} \log 2\pi - \frac{1}{2} \mathbb{E}_{q_{\Lambda_k\mu_k}}^{(i)} \{ (\mathbf{x}_n - \mu_k)^\top \Lambda_k (\mathbf{x}_n - \mu_k) \}
\]
Derivations for the Bayesian mixture of Gaussian

**E-Z-Step:**

\[
\log q_Z^{(i)}(Z) = \mathbb{E}_{q_{\Lambda,\mu,\pi}}^{(i)} \{ \log p(Z|X, \Lambda, \mu, \pi) \} + \text{const.}
\]

\[
= \sum_{n,k=1}^{N,K} \mathbb{I}\{Z_n = k\} \log \rho_{n,k}^{(i)} + \text{const.}
\]

where

\[
\rho_{n,k}^{(i)} = \mathbb{E}_{q_{\pi}}^{(i)} \{ \log \pi_k \} + \frac{1}{2} \mathbb{E}_{q_{\Lambda_k}}^{(i)} \{ \log |\Lambda_k| \} - \frac{D}{2} \log 2\pi - \frac{1}{2} \mathbb{E}_{q_{\Lambda_k,\mu_k}}^{(i)} \{(x_n - \mu_k)^\top \Lambda_k (x_n - \mu_k)\}
\]

It follows that

\[
q_Z^{(i)}(Z) = \prod_{n=1}^{N} q_{Z_n}^{(i)}(k)
\]

where

\[
q_{Z_n}^{(i)}(k) = \frac{\rho_{n,k}^{(i)}}{\sum_{j=1}^{K} \rho_{n,j}^{(i)}} = r_{n,k}^{(i)}
\]
Derivations for the Bayesian mixture of Gaussian

**E-Z-Step:**

\[
\log q_Z^{(i)}(Z) = \mathbb{E}_{q_{\Lambda \mu \pi}} \left\{ \log p(Z | X, \Lambda, \mu, \pi) \right\} + \text{const.}
\]

\[
= \sum_{n,k=1}^{N,K} \mathbb{I}\{Z_n = k\} \log \rho_{n,k}^{(i)} + \text{const.}
\]

where

\[
\rho_{n,k}^{(i)} = \mathbb{E}_{q_{\pi}} \left\{ \log \pi_k \right\} + \frac{1}{2} \mathbb{E}_{q_{\Lambda_k}} \left\{ \log |\Lambda_k| \right\} - \frac{D}{2} \log 2\pi - \frac{1}{2} \mathbb{E}_{q_{\Lambda_k \mu_k}} \left\{ (x_n - \mu_k)^\top \Lambda_k (x_n - \mu_k) \right\}
\]

It follows that

\[
q_Z^{(i)}(Z) = \prod_{n=1}^{N} q_{Z_n}(k) \quad \text{where} \quad q_{Z_n}(k) = \frac{\rho_{n,k}^{(i)}}{\sum_{j=1}^{K} \rho_{n,j}^{(i)}} = r_{n,k}^{(i)}
\]

Finally, we can express \( r_{n,k}^{(i)} \) as a function of the parameters calculated in previous step:

\[
r_{n,k}^{(i)} \propto |W_k^{(i)}|^{1/2} \exp \left( \psi(\alpha_0 + N_k^{(i)}) + \sum_{i=1}^{D} \psi \left( \frac{\nu_k^{(i)} + 1 - i}{2} \right) - \frac{D}{2 N_k^{(i)} + 2} - \frac{\nu_k^{(i)}}{2} (x_n - m_k^{(i)})^\top W_k^{(i)} (x_n - m_k^{(i)}) \right)
\]

where \( \psi(.) \) denotes the digamma function.
Inference

VB-EM in action

• Initialization: Random means + GMM E-step for $q_Z^{(0)}(Z)$

• E-$\Lambda\mu\pi$-Step: $q_{\Lambda\mu\pi}^{(i)}(\Lambda, \mu, \pi) \propto \exp\left(\mathbb{E}_{q_{Z}^{(i-1)}}\{\log p(\Lambda, \mu, \pi | X, Z)\}\right)$

• E-Z-Step: $q_Z^{(i)}(Z) \propto \exp\left(\mathbb{E}_{q_{\Lambda\mu\pi}^{(i)}}\{\log p(Z | X, \Lambda, \mu, \pi)\}\right)$

• Convergence

$K = 5, \alpha_0 = 0.01$
Inference

VB-EM in action

- **Initialization**: Random means + GMM E-step for $q_Z^{(0)}(Z)$

- **E-Λμπ-Step**: $q_{Λμπ}^{(i)}(Λ, μ, π) \propto \exp\left(\mathbb{E}_{q_Z^{(i-1)}}\{\log p(Λ, μ, π | X, Z)\}\right)$

- **E-Z-Step**: $q_Z^{(i)}(Z) \propto \exp\left(\mathbb{E}_{q_{Λμπ}^{(i)}}\{\log p(Z | X, Λ, μ, π)\}\right)$

- **Convergence**
**VB-EM in action**

**Initialization:** Random means + GMM E-step for $q_Z^{(0)}(Z)$

- **E-$\Lambda\mu\pi$-Step:** $q_{\Lambda\mu\pi}^{(i)}(\Lambda, \mu, \pi) \propto \exp \left( \mathbb{E}_{q_Z^{(i-1)}} \{ \log p(\Lambda, \mu, \pi | X, Z) \} \right)$

- **E-Z-Step:** $q_Z^{(i)}(Z) \propto \exp \left( \mathbb{E}_{q_{\Lambda\mu\pi}^{(i)}} \{ \log p(Z | X, \Lambda, \mu, \pi) \} \right)$

- **Convergence**
Inference

VB-EM in action

- **Initialization**: Random means + GMM E-step for $q_Z^{(0)}(Z)$
- **E-$\Lambda\mu\pi$-Step**: $q_{\Lambda\mu\pi}^{(i)}(\Lambda, \mu, \pi) \propto \exp\left( \mathbb{E}_{q_Z^{(i-1)}} \{ \log p(\Lambda, \mu, \pi | X, Z) \} \right)$
- **E-Z-Step**: $q_Z^{(i)}(Z) \propto \exp\left( \mathbb{E}_{q_{\Lambda\mu\pi}^{(i)}} \{ \log p(Z | X, \Lambda, \mu, \pi) \} \right)$
- **Convergence**

$K = 5, \alpha_0 = 0.01$
VB-EM in action

\[
\begin{align*}
\text{\textbullet} \quad \text{Initialization: Random means + GMM E-step for } q^{(0)}_Z(Z) \\
\text{\textbullet} \quad \text{E-\(\Lambda\mu\pi\)-Step: } q^{(i)}_{\Lambda\mu\pi}(\Lambda, \mu, \pi) \propto \exp \left( \mathbb{E}_{q^{(i-1)}_Z} \{ \log p(\Lambda, \mu, \pi | X, Z) \} \right) \\
\text{\textbullet} \quad \text{E-Z-Step: } q^{(i)}_Z(Z) \propto \exp \left( \mathbb{E}_{q^{(i)}_{\Lambda\mu\pi}} \{ \log p(Z | X, \Lambda, \mu, \pi) \} \right) \\
\text{\textbullet} \quad \text{Convergence}
\end{align*}
\]

\[K = 5, \ \alpha_0 = 0.01\]
Inference

VB-EM in action

- **Initialization**: Random means + GMM E-step for $q^{(0)}_Z(Z)$
- **E-$\Lambda\mu\pi$-Step**: $q^{(i)}_{\Lambda\mu\pi}(\Lambda, \mu, \pi) \propto \exp \left( \mathbb{E}_{q^{(i-1)}_Z}\{\log p(\Lambda, \mu, \pi|X, Z)\} \right)$
- **E-$Z$-Step**: $q^{(i)}_Z(Z) \propto \exp \left( \mathbb{E}_{q^{(i)}_{\Lambda\mu\pi}}\{\log p(Z|X, \Lambda, \mu, \pi)\} \right)$
- **Convergence**
Bayesian Inference: Examples

Inference

VB-EM in action

- Initialization: Random means + GMM E-step for $q^{(0)}_Z(Z)$
- E-Λμπ-Step: $q^{(i)}_{Λμπ}(Λ, μ, π) \propto \exp \left( \mathbb{E}_{q^{(i-1)}_Z} \{ \log p(Λ, μ, π | X, Z) \} \right)$
- E-Z-Step: $q^{(i)}_Z(Z) \propto \exp \left( \mathbb{E}_{q^{(i)}_{Λμπ}} \{ \log p(Z | X, Λ, μ, π) \} \right)$
- Convergence

$K = 5, \ a_0 = 0.01$
Bayesian Inference: Examples

VB-EM in action

- Initialization: Random means + GMM E-step for $q_Z^{(0)}(Z)$
- E-$\Lambda\mu\pi$-Step: $q^{(i)}_{\Lambda\mu\pi}(\Lambda, \mu, \pi) \propto \exp \left( \mathbb{E}_{q_Z^{(i-1)}} \{ \log p(\Lambda, \mu, \pi | X, Z) \} \right)$
- E-$Z$-Step: $q^{(i)}_Z(Z) \propto \exp \left( \mathbb{E}_{q^{(i)}_{\Lambda\mu\pi}} \{ \log p(Z | X, \Lambda, \mu, \pi) \} \right)$
- Convergence

$K = 5, \alpha_0 = 0.01$
Bayesian Inference: Examples

Inference

VB-EM in action

• Initialization: Random means + GMM E-step for $q_Z^{(0)}(Z)$

• $E$-Λμπ-Step: $q_{Λμπ}^{(i)}(Λ, μ, π) \propto \exp \left( \mathbb{E}_{q_Z^{(i-1)}} \{ \log p(Λ, μ, π|X, Z) \} \right)$

• $E$-Z-Step: $q_Z^{(i)}(Z) \propto \exp \left( \mathbb{E}_{q_{Λμπ}^{(i)}} \{ \log p(Z|X, Λ, μ, π) \} \right)$

• Convergence

$K = 5, \ \alpha_0 = 0.01$
VB-EM in action

• **Initialization**: Random means + GMM E-step for $q_{Z}^{(0)}(Z)$

• **E-Λμπ-Step**:
  $$q_{Λμπ}^{(i)}(Λ, μ, π) \propto \exp \left( \mathbb{E}_{q_{Z}^{(i-1)}} \{ \log p(Λ, μ, π | X, Z) \} \right)$$

• **E-Z-Step**:
  $$q_{Z}^{(i)}(Z) \propto \exp \left( \mathbb{E}_{q_{Λμπ}^{(i)}} \{ \log p(Z | X, Λ, μ, π) \} \right)$$

• **Convergence**

$K = 5$, $α_0 = 0.01$
**Inference**

**VB-EM in action**

- **Initialization:** Random means + GMM E-step for $q^{(0)}_Z(Z)$

- **E-Λμπ-Step:** $q^{(i)}_{Λμπ}(Λ, μ, π) \propto \exp \left( \mathbb{E}_{q^{(i-1)}_Z} \{ \log p(Λ, μ, π | X, Z) \} \right)$

- **E-Z-Step:** $q^{(i)}_Z(Z) \propto \exp \left( \mathbb{E}_{q^{(i)}_{Λμπ}} \{ \log p(Z | X, Λ, μ, π) \} \right)$

- **Convergence**

$K = 5, \, \alpha_0 = 0.01$
Bayesian Inference: Examples

Variational methods

Inference

VB-EM in action

- Initialization: Random means + GMM E-step for \( q_Z^{(0)}(Z) \)
- E-\( \Lambda \mu \pi \)-Step: \( q_{\Lambda \mu \pi}^{(i)}(\Lambda, \mu, \pi) \propto \exp \left( \mathbb{E}_{q_{Z}^{(i-1)}} \{ \log p(\Lambda, \mu, \pi | X, Z) \} \right) \)
- E-Z-Step: \( q_{Z}^{(i)}(Z) \propto \exp \left( \mathbb{E}_{q_{\Lambda \mu \pi}^{(i)}} \{ \log p(Z | X, \Lambda, \mu, \pi) \} \right) \)
- Convergence

\[ K = 5, \quad \alpha_0 = 0.01 \]
Inference

VB-EM in action

- **Initialization**: Random means + GMM E-step for $q_Z^{(0)}(Z)$

- **E-$\Lambda\mu\pi$-Step**: $q_{\Lambda\mu\pi}^{(i)}(\Lambda, \mu, \pi) \propto \exp \left( \mathbb{E}_{q_{Z}^{(i-1)}} \{ \log p(\Lambda, \mu, \pi | X, Z) \} \right)$

- **E-Z-Step**: $q_Z^{(i)}(Z) \propto \exp \left( \mathbb{E}_{q_{\Lambda\mu\pi}^{(i)}} \{ \log p(Z | X, \Lambda, \mu, \pi) \} \right)$

- **Convergence**
VB-EM in action

\( q_{\Lambda \mu \pi}^{(i)}(\Lambda, \mu, \pi) \propto \exp \left( \mathbb{E}_{q_{Z}^{(i-1)}} \{ \log p(\Lambda, \mu, \pi | X, Z) \} \right) \)

\( q_{Z}^{(i)}(Z) \propto \exp \left( \mathbb{E}_{q_{\Lambda \mu \pi}^{(i)}} \{ \log p(Z | X, \Lambda, \mu, \pi) \} \right) \)

\( K = 5, \alpha_0 = 0.01 \)
Inference

VB-EM in action

- Initialization: Random means + GMM E-step for $q_Z^{(0)}(Z)$
- **E-$\Lambda\mu\pi$-Step**: $q^{(i)}_{\Lambda\mu\pi}(\Lambda, \mu, \pi) \propto \exp \left( \mathbb{E}_{q^{(i-1)}_Z} \{ \log p(\Lambda, \mu, \pi | X, Z) \} \right)$
- **E-$Z$-Step**: $q^{(i)}_Z(Z) \propto \exp \left( \mathbb{E}_{q^{(i)}_{\Lambda\mu\pi}} \{ \log p(Z | X, \Lambda, \mu, \pi) \} \right)$
- Convergence

$K = 5$, $\alpha_0 = 0.01$
Bayesian Inference: Examples

Inference

VB-EM in action

- **Initialization**: Random means + GMM E-step for $q_Z^{(0)}(Z)$

- **E-Λμπ-Step**: $q_{Λμπ}^{(i)}(Λ, μ, π) \propto \exp \left( \mathbb{E}_{q_Z^{(i-1)}} \{ \log p(Λ, μ, π | X, Z) \} \right)$

- **E-Z-Step**: $q_Z^{(i)}(Z) \propto \exp \left( \mathbb{E}_{q_{Λμπ}^{(i)}} \{ \log p(Z | X, Λ, μ, π) \} \right)$

- **Convergence**

$K = 5, \alpha_0 = 0.01$
Inference

VB-EM in action

- **Initialization:** Random means + GMM E-step for $q_{Z}^{(0)}(Z)$
- **E-Λμπ-Step:** $q_{λμπ}^{(i)}(Λ, μ, π) \propto \exp \left( \mathbb{E}_{q_{Z}^{(i-1)}} \{ \log p(Λ, μ, π | X, Z) \} \right)$
- **E-Z-Step:** $q_{Z}^{(i)}(Z) \propto \exp \left( \mathbb{E}_{q_{λμπ}^{(i)}} \{ \log p(Z | X, Λ, μ, π) \} \right)$
- **Convergence**

$K = 5, \; α_0 = 0.01$
Bayesian Inference: Examples

**Inference**

**VB-EM in action**

- **Initialization:** Random means + GMM E-step for $q^{(0)}_Z(Z)$
- **E-Λμπ-Step:**
  
  $$q^{(i)}_{Λμπ}(Λ, μ, π) \propto \exp \left( \mathbb{E}_{q^{(i-1)}_Z} \{ \log p(Λ, μ, π | X, Z) \} \right)$$

- **E-Z-Step:**
  
  $$q^{(i)}_Z(Z) \propto \exp \left( \mathbb{E}_{q^{(i)}_{Λμπ}} \{ \log p(Z | X, Λ, μ, π) \} \right)$$

- **Convergence**


\[ K = 5, \quad α_0 = 0.01 \]
Inference

VB-EM in action

Inference

• Initialization: Random means + GMM E-step for $q_{Z}^{(0)}(Z)$

• E-Λμπ-Step : $q_{Λμπ}^{(i)}(Λ, μ, π) \propto \exp \left( \mathbb{E}_{q_{Z}^{(i-1)}} \{ \log p(Λ, μ, π | X, Z) \} \right)$

• E-Z-Step: $q_{Z}^{(i)}(Z) \propto \exp \left( \mathbb{E}_{q_{Λμπ}^{(i)}} \{ \log p(Z | X, Λ, μ, π) \} \right)$

• Convergence

$K = 5, \alpha_0 = 0.01$
Inference

VB-EM in action

- Initialization: Random means + GMM E-step for $q^{(0)}_Z(Z)$

- E-Λμπ-Step: $q^{(i)}_{Λμπ}(Λ, μ, π) \propto \exp \left( \mathbb{E}_{q^{(i-1)}_Z} \{ \log p(Λ, μ, π | X, Z) \} \right)$

- E-Z-Step: $q^{(i)}_Z(Z) \propto \exp \left( \mathbb{E}_{q^{(i)}_{Λμπ}} \{ \log p(Z | X, Λ, μ, π) \} \right)$

- Convergence

$K = 5, \; α_0 = 0.01$
Inference

VB-EM in action

- **Initialization:** Random means + GMM E-step for $q^{(0)}_Z(Z)$
- **E-Λμπ-Step:** $q^{(i)}_{Λμπ}(Λ, μ, π) \propto \exp \left( \mathbb{E}_{q^{(i-1)}_Z} \{ \log p(Λ, μ, π|X, Z) \} \right)$
- **E-Z-Step:** $q^{(i)}_Z(Z) \propto \exp \left( \mathbb{E}_{q^{(i)}_{Λμπ}} \{ \log p(Z|X, Λ, μ, π) \} \right)$
- **Convergence**

$k = 5, \alpha_0 = 0.01$
Bayesian Inference: Examples

Inference

VB-EM in action

• Initialization: Random means + GMM E-step for $q_Z^{(0)}(Z)$

• E-Λμπ-Step: $q_{Λμπ}^{(i)}(Λ, μ, π) \propto \exp \left( \mathbb{E}_{q_Z^{(i-1)}} \{ \log p(Λ, μ, π | X, Z) \} \right)$

• E-Z-Step: $q_Z^{(i)}(Z) \propto \exp \left( \mathbb{E}_{q_{Λμπ}^{(i)}} \{ \log p(Z | X, Λ, μ, π) \} \right)$

• Convergence

$K = 5, \alpha_0 = 0.01$
Bayesian Inference: Examples

VB-EM in action

• Initialization: Random means + GMM E-step for $q_Z^{(0)}(Z)$

• E-Λμπ-Step: $q_{Λμπ}^{(i)}(Λ, μ, π) \propto \exp \left( \mathbb{E}_{q_Z^{(i-1)}} \{ \log p(Λ, μ, π | X, Z) \} \right)$

• E-Z-Step: $q_Z^{(i)}(Z) \propto \exp \left( \mathbb{E}_{q_{Λμπ}^{(i)}} \{ \log p(Z | X, Λ, μ, π) \} \right)$

• Convergence

Conclusions on GMM VB-EM

• Similar computational time as GMM-EM (though slightly more iterations)

• Priors on Gaussian weights handle automatically degenerate or unused clusters

• Determination of $K$

• Works even for very small data samples

$K = 5$, $α_0 = 0.01$
• What is Bayesian inference?
  – Overview
  – Classical vs. Bayesian approach
  – Bayes Theorem & Example
  – General Methodology

• Bayesian inference by examples
  – Direct inference
  – The Expectation-Maximization algorithm
  – Variational Bayes methods

• Markov chain Monte Carlo
• What is Bayesian inference?
  – Overview
  – Classical vs. Bayesian approach
  – Bayes Theorem & Example
  – General Methodology
• Bayesian inference by examples
  – Direct inference
  – The Expectation-Maximization algorithm
  – Variational Bayes methods
• Markov chain Monte Carlo