Модели со скрытыми переменными. EM-алгоритм

Евгений Бурнаев

Сколтех, Москва, Россия

Skoltech

Skolkovo Institute of Science and Technology

План

- Мера расхождения между двумя распределениями, определенными в одних и тех же областях

$$KL(q\|p) = -\int q(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} =$$

- Информационно-теоретическая интерпретация

$$\mathrm{KL} = \mathrm{Cross\,Entropy} - \mathrm{Entropy}$$

- Если мы минимизируем KL по $q(\cdot)$ приближение должно быть хорошим там, где $q(\cdot)$ имеет большие значения

- Мера расхождения между двумя распределениями, определенными в одних и тех же областях

$$KL(q\|p) = -\int q(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} =$$

$$= \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x} = \mathbb{E}_{\mathbf{x}\sim q(\cdot)} \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \geq 0$$

- Информационно-теоретическая интерпретация

$$\mathrm{KL} = \mathrm{Cross\,Entropy} - \mathrm{Entropy}$$

- Если мы минимизируем KL по $q(\cdot)$ приближение должно быть хорошим там, где $q(\cdot)$ имеет большие значения

- Мера расхождения между двумя распределениями, определенными в одних и тех же областях

$$KL(q\|p) = -\int q(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} =$$

$$= \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x} = \mathbb{E}_{\mathbf{x}\sim q(\cdot)} \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \geq 0$$

- Информационно-теоретическая интерпретация

$$\mathrm{KL} = \mathrm{Cross\,Entropy} - \mathrm{Entropy}$$

- Если мы минимизируем KL по $q(\cdot)$ приближение должно быть хорошим там, где $q(\cdot)$ имеет большие значения

- Мера расхождения между двумя распределениями, определенными в одних и тех же областях

$$KL(q\|p) = -\int q(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} =$$

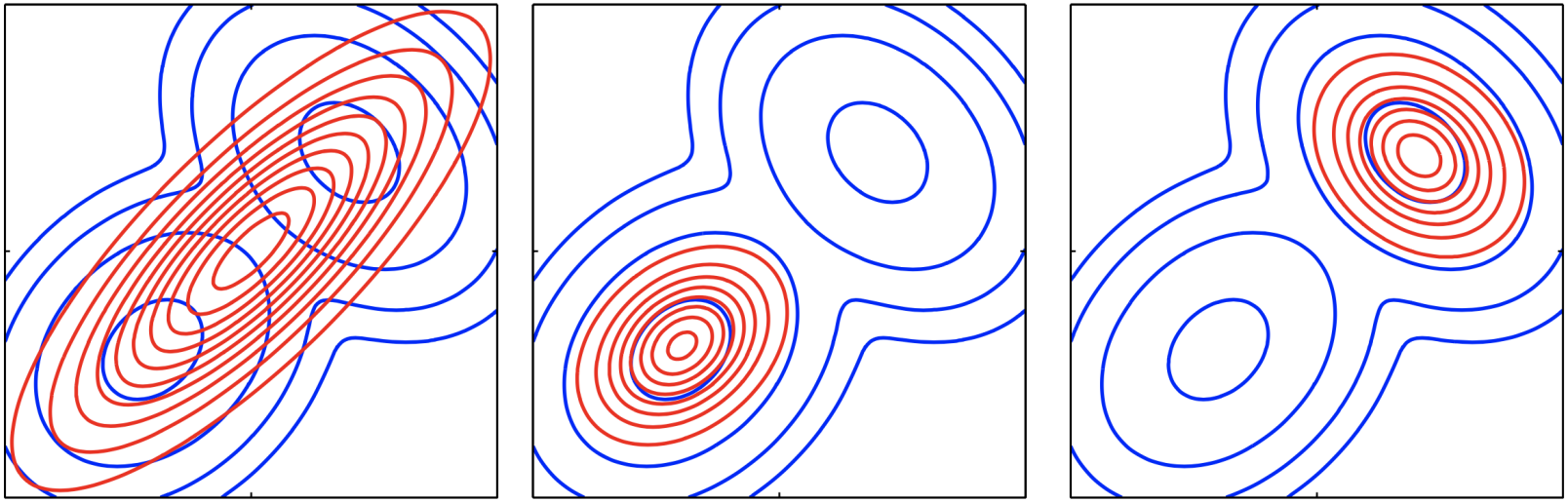
$$= \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x} = \mathbb{E}_{\mathbf{x}\sim q(\cdot)} \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \geq 0$$

- Информационно-теоретическая интерпретация

$$\mathrm{KL} = \mathrm{Cross\,Entropy} - \mathrm{Entropy}$$

- Если мы минимизируем KL по $q(\cdot)$ приближение должно быть хорошим там, где $q(\cdot)$ имеет большие значения

Расстояние Кульбака-Лейблера

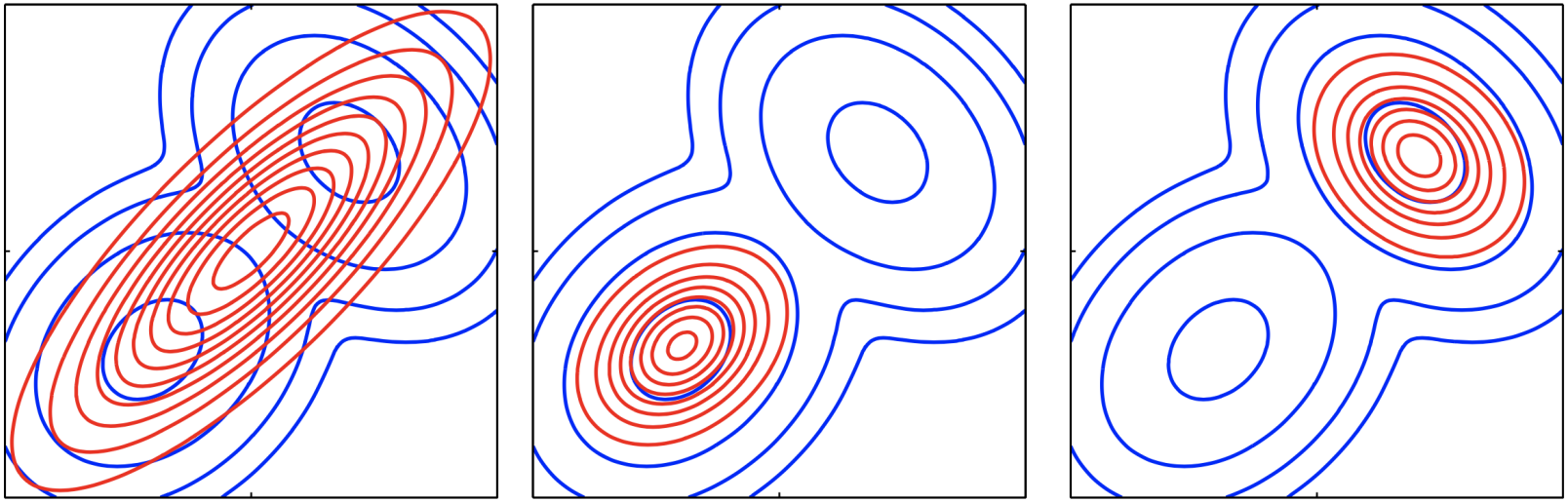


(a) Синие контуры: бимодальное распределение, представляющее смесь двух Гауссовских распределений $p(\mathbf{z})$. Красные контуры: одиночное Гауссовское распределение $q(\mathbf{z})$ которое наилучшим образом приближает $p(\mathbf{z})$ к минимуму $KL(p\|q)$

(b) Как и в (a), но теперь $q(\mathbf{z})$ находится путем численной минимизации $KL(q\|p)$

(c) Как в (b), но показывает другой локальный минимум $KL(q\|p)$

Расстояние Кульбака-Лейблера

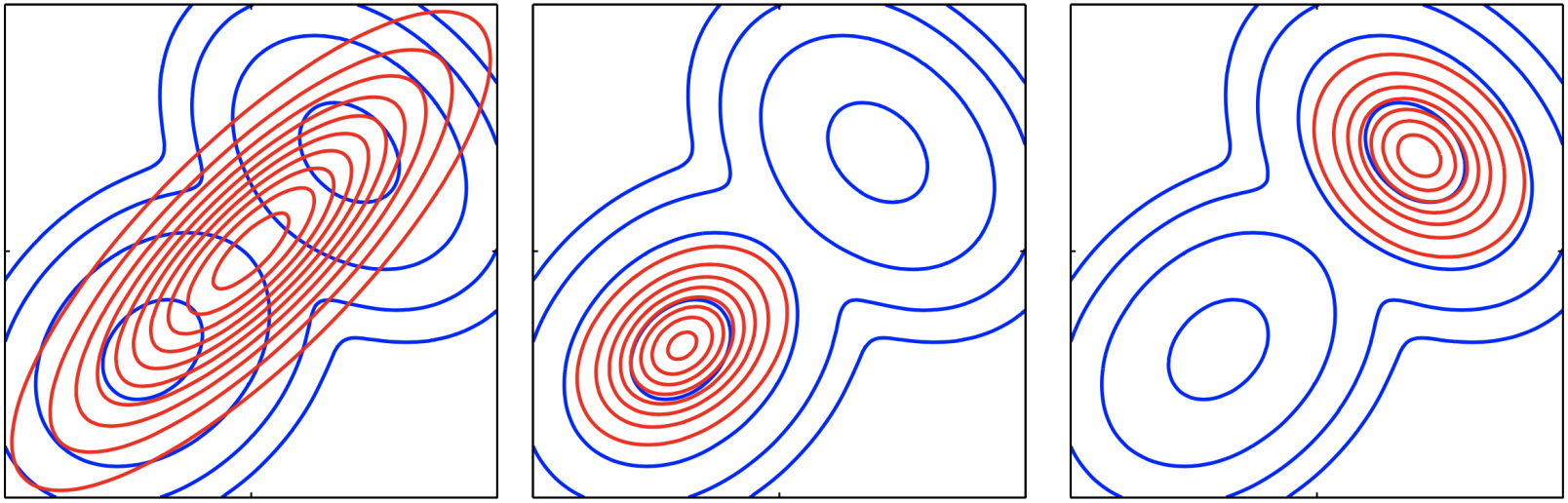


(a) Синие контуры: бимодальное распределение, представляющее смесь двух Гауссовских распределений $p(\mathbf{z})$. Красные контуры: одиночное Гауссовское распределение $q(\mathbf{z})$ которое наилучшим образом приближает $p(\mathbf{z})$ к минимуму $KL(p\|q)$

(b) Как и в (a), но теперь $q(\mathbf{z})$ находится путем численной минимизации $KL(q\|p)$

(c) Как в (b), но показывает другой локальный минимум $KL(q\|p)$

Расстояние Кульбака-Лейблера



(a) Синие контуры: бимодальное распределение, представляющее смесь двух Гауссовских распределений $p(\mathbf{z})$. Красные контуры: одиночное Гауссовское распределение $q(\mathbf{z})$ которое наилучшим образом приближает $p(\mathbf{z})$ к минимуму $KL(p\|q)$

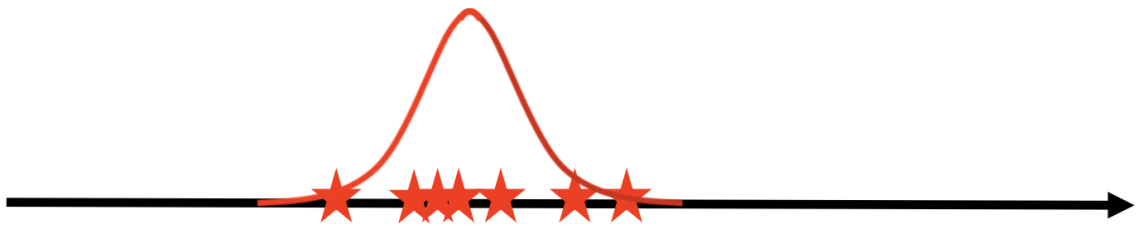
(b) Как и в (a), но теперь $q(\mathbf{z})$ находится путем численной минимизации $KL(q\|p)$

(c) Как в (b), но показывает другой локальный минимум $KL(q\|p)$

Моделирование скрытых переменных: Пример

- У нас есть набор точек, сгенерированный из Гауссовского распределения

$$x_i \sim \mathcal{N}(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

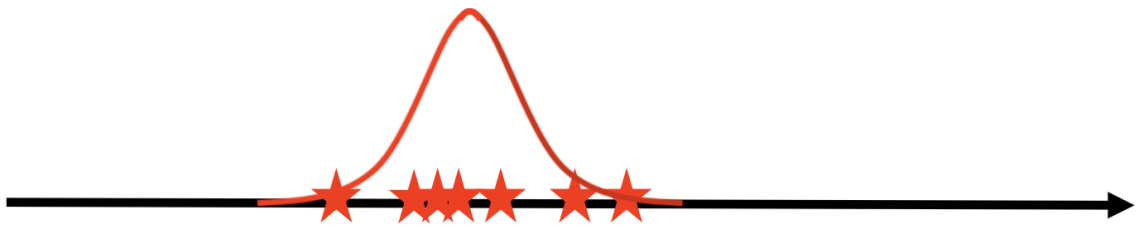


- Оценим его параметры μ и σ: выборочное среднее и дисперсия

Моделирование скрытых переменных: Пример

- У нас есть набор точек, сгенерированный из Гауссовского распределения

$$x_i \sim \mathcal{N}(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$



- Оценим его параметры μ и σ: выборочное среднее и дисперсия

Моделирование скрытых переменных: Пример

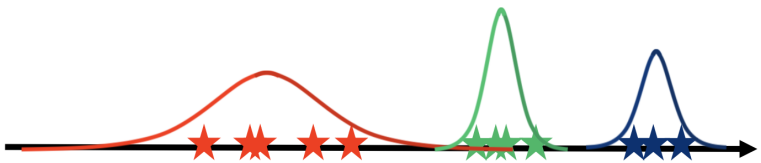
- Несколько наборов точек от разных гауссианов

- Мы должны оценить параметры этих гауссиан и их веса
- Задача простая, если мы знаем, из какой гауссианы сгенерирован каждый объект
- Использование единичной гауссовой модели приводит к неточным результатам

Моделирование скрытых переменных: Пример

- Несколько наборов точек от разных гауссианов

- Мы должны оценить параметры этих гауссиан и их веса

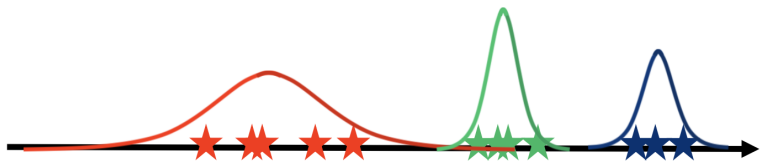


- Задача простая, если мы знаем, из какой гауссианы сгенерирован каждый объект
- Использование единичной гауссовой модели приводит к неточным результатам

Моделирование скрытых переменных: Пример

- Несколько наборов точек от разных гауссианов

- Мы должны оценить параметры этих гауссиан и их веса

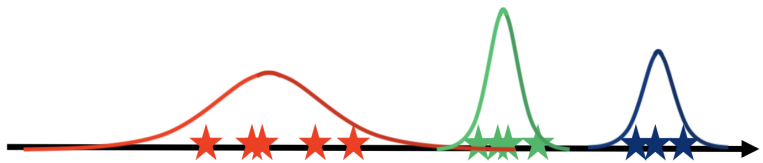


- Задача простая, если мы знаем, из какой гауссианы сгенерирован каждый объект
- Использование единичной гауссовой модели приводит к неточным результатам

Моделирование скрытых переменных: Пример

- Несколько наборов точек от разных гауссианов

- Мы должны оценить параметры этих гауссиан и их веса



- Задача простая, если мы знаем, из какой гауссианы сгенерирован каждый объект
- Использование единичной гауссовой модели приводит к неточным результатам

Гауссовская смесь

- Для каждого x_i мы вводим дополнительный z_i, обозначающий индекс Гауссовского распределения, из которого был сгенерирован i-й объект
- Модель выглядит следующим образом:

- Здесь $\pi_j = p(z_i = j)$ априорная вероятность j-го Гауссовского распределения, а $\boldsymbol{\theta} = \{\mu_j, \sigma_j, \pi_j\}_{j=1}^K$ параметры, подлежащие оценке
- Если мы знаем $\mathbf{X}$ и $\mathbf{Z}$, можно использовать метод максимального правдоподобия (MLE).

$$\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

Гауссовская смесь

- Для каждого x_i мы вводим дополнительный z_i, обозначающий индекс Гауссовского распределения, из которого был сгенерирован i-й объект
- Модель выглядит следующим образом:

- Здесь $\pi_j = p(z_i = j)$ априорная вероятность j-го Гауссовского распределения, а $\boldsymbol{\theta} = \{\mu_j, \sigma_j, \pi_j\}_{j=1}^{K}$ параметры, подлежащие оценке
- Если мы знаем $\mathbf{X}$ и $\mathbf{Z}$, можно использовать метод максимального правдоподобия (MLE).

$$\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

Гауссовская смесь

- Для каждого x_i мы вводим дополнительный z_i, обозначающий индекс Гауссовского распределения, из которого был сгенерирован i-й объект
- Модель выглядит следующим образом:

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \prod_{i=1}^{m} p(x_i, z_i|\boldsymbol{\theta}) =$$

- Здесь $\pi_j = p(z_i = j)$ априорная вероятность j-го Гауссовского распределения, а $\boldsymbol{\theta} = \{\mu_j, \sigma_j, \pi_j\}_{j=1}^{K}$ параметры, подлежащие оценке
- Если мы знаем $\mathbf{X}$ и $\mathbf{Z}$, можно использовать метод максимального правдоподобия (MLE).

$$\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

Гауссовская смесь

- Для каждого x_i мы вводим дополнительный z_i, обозначающий индекс Гауссовского распределения, из которого был сгенерирован i-й объект
- Модель выглядит следующим образом:

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \prod_{i=1}^{m} p(x_i, z_i|\boldsymbol{\theta}) = \prod_{i=1}^{m} p(x_i|z_i, \boldsymbol{\theta}) p(z_i|\boldsymbol{\theta})$$

- Здесь $\pi_j = p(z_i = j)$ априорная вероятность j-го Гауссовского распределения, а $\boldsymbol{\theta} = \{\mu_j, \sigma_j, \pi_j\}_{j=1}^{K}$ параметры, подлежащие оценке
- Если мы знаем $\mathbf{X}$ и $\mathbf{Z}$, можно использовать метод максимального правдоподобия (MLE).

$$\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

Гауссовская смесь

- Для каждого x_i мы вводим дополнительный z_i, обозначающий индекс Гауссовского распределения, из которого был сгенерирован i-й объект
- Модель выглядит следующим образом:

$$
\begin{aligned}
p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) &= \prod_{i=1}^{m} p(x_i, z_i|\boldsymbol{\theta}) = \\
&= \prod_{i=1}^{m} p(x_i|z_i, \boldsymbol{\theta}) p(z_i|\boldsymbol{\theta}) \\
&= \prod_{i=1}^{m} \pi_{z_i} \mathcal{N}(x_i|\mu_{z_i}, \sigma_{z_i}^2)
\end{aligned}
$$

- Здесь $\pi_j = p(z_i = j)$ априорная вероятность j-го Гауссовского распределения, а $\boldsymbol{\theta} = \{\mu_j, \sigma_j, \pi_j\}_{j=1}^{K}$ параметры, подлежащие оценке
- Если мы знаем $\mathbf{X}$ и $\mathbf{Z}$, можно использовать метод максимального правдоподобия (MLE).

$$\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

Гауссовская смесь

- Для каждого x_i мы вводим дополнительный z_i, обозначающий индекс Гауссовского распределения, из которого был сгенерирован i-й объект
- Модель выглядит следующим образом:

$$
\begin{aligned}
p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) &= \prod_{i=1}^{m} p(x_i, z_i|\boldsymbol{\theta}) = \\
&= \prod_{i=1}^{m} p(x_i|z_i, \boldsymbol{\theta}) p(z_i|\boldsymbol{\theta}) \\
&= \prod_{i=1}^{m} \pi_{z_i} \mathcal{N}(x_i|\mu_{z_i}, \sigma_{z_i}^2)
\end{aligned}
$$

- Здесь $\pi_j = p(z_i = j)$ априорная вероятность j-го Гауссовского распределения, а $\boldsymbol{\theta} = \{\mu_j, \sigma_j, \pi_j\}_{j=1}^{K}$ параметры, подлежащие оценке
- Если мы знаем $\mathbf{X}$ и $\mathbf{Z}$, можно использовать метод максимального правдоподобия (MLE).

$$\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

Гауссовская смесь

- Для каждого x_i мы вводим дополнительный z_i, обозначающий индекс Гауссовского распределения, из которого был сгенерирован i-й объект
- Модель выглядит следующим образом:

$$
\begin{aligned}
p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) &= \prod_{i=1}^{m} p(x_i, z_i|\boldsymbol{\theta}) = \\
&= \prod_{i=1}^{m} p(x_i|z_i, \boldsymbol{\theta}) p(z_i|\boldsymbol{\theta}) \\
&= \prod_{i=1}^{m} \pi_{z_i} \mathcal{N}(x_i|\mu_{z_i}, \sigma_{z_i}^2)
\end{aligned}
$$

- Здесь $\pi_j = p(z_i = j)$ априорная вероятность j-го Гауссовского распределения, а $\boldsymbol{\theta} = \{\mu_j, \sigma_j, \pi_j\}_{j=1}^{K}$ параметры, подлежащие оценке
- Если мы знаем $\mathbf{X}$ и $\mathbf{Z}$, можно использовать метод максимального правдоподобия (MLE).

$$
\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})
$$

Гауссовская смесь

- Мы не знаем $\mathbf{Z} \Rightarrow$ мы максимизируем по $\boldsymbol{\theta}$ логарифм маргинального правдоподобия

$$\log p(\mathbf{X}|\boldsymbol{\theta})$$

- Для любого распределения $q(\mathbf{Z})$ получаем

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{Z}$$

- Поскольку $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \cdot p(\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})$, получаем

Гауссовская смесь

- Мы не знаем $\mathbf{Z} \Rightarrow$ мы максимизируем по $\boldsymbol{\theta}$ логарифм маргинального правдоподобия

$$\log p(\mathbf{X}|\boldsymbol{\theta})$$

- Для любого распределения $q(\mathbf{Z})$ получаем

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{Z}$$

- Поскольку $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \cdot p(\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})$, получаем

Гауссовская смесь

- Мы не знаем $\mathbf{Z} \Rightarrow$ мы максимизируем по $\boldsymbol{\theta}$ логарифм маргинального правдоподобия

$$\log p(\mathbf{X}|\boldsymbol{\theta})$$

- Для любого распределения $q(\mathbf{Z})$ получаем

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{Z}$$

- Поскольку $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \cdot p(\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})$, получаем

Гауссовская смесь

- Мы не знаем $\mathbf{Z} \Rightarrow$ мы максимизируем по $\boldsymbol{\theta}$ логарифм маргинального правдоподобия

$$\log p(\mathbf{X}|\boldsymbol{\theta})$$

- Для любого распределения $q(\mathbf{Z})$ получаем

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{Z}$$

- Поскольку $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \cdot p(\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})$, получаем

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{Z} =$$

Гауссовская смесь

- Мы не знаем $\mathbf{Z} \Rightarrow$ мы максимизируем по $\boldsymbol{\theta}$ логарифм маргинального правдоподобия

$$\log p(\mathbf{X}|\boldsymbol{\theta})$$

- Для любого распределения $q(\mathbf{Z})$ получаем

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{Z}$$

- Поскольку $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \cdot p(\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})$, получаем

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{Z} =$$

$$= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} d\mathbf{Z}$$

Гауссовская смесь

- Мы не знаем $\mathbf{Z} \Rightarrow$ мы максимизируем по $\boldsymbol{\theta}$ логарифм маргинального правдоподобия

$$\log p(\mathbf{X}|\boldsymbol{\theta})$$

- Для любого распределения $q(\mathbf{Z})$ получаем

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{Z}$$

- Поскольку $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \cdot p(\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})$, получаем

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{Z} =$$

$$= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} d\mathbf{Z} = \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z}) p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z}) p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} d\mathbf{Z}$$

- Мы не знаем $\mathbf{Z} \Rightarrow$ мы максимизируем по $\boldsymbol{\theta}$ логарифм маргинального правдоподобия

$$\log p(\mathbf{X}|\boldsymbol{\theta})$$

- Для любого распределения $q(\mathbf{Z})$ получаем

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{Z}$$

- Поскольку $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \cdot p(\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})$, получаем

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{Z} =$$

$$= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} d\mathbf{Z} = \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z}) p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z}) p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} d\mathbf{Z}$$

$$= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} + \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} d\mathbf{Z}$$

- Мы получаем

$$\log p(\mathbf{X}|\boldsymbol{\theta}) =$$

$$= \underbrace{\int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z}}_{\text{Нижняя оценка обусловленности ELBO } \mathcal{L}(q, \boldsymbol{\theta})} + \underbrace{\int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} d\mathbf{Z}}_{\text{Неотрицательное}}$$

- Таким образом

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p) \geq \mathcal{L}(q, \boldsymbol{\theta})$$

- Вместо оптимизации $\log p(\mathbf{X}|\boldsymbol{\theta})$ мы оптимизируем ELBO $\mathcal{L}(q, \boldsymbol{\theta})$ как по $\boldsymbol{\theta}$, так и по $q(\mathbf{Z})$
- Алгоритм блочно-координатного спуска известен как EM-алгоритм

- Мы получаем

$$\log p(\mathbf{X}|\boldsymbol{\theta}) =$$

$$= \underbrace{\int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z}}_{\text{Нижняя оценка обусловленности ELBO } \mathcal{L}(q, \boldsymbol{\theta})} + \underbrace{\int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} d\mathbf{Z}}_{\text{Неотрицательное}}$$

- Таким образом

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p) \geq \mathcal{L}(q, \boldsymbol{\theta})$$

- Вместо оптимизации $\log p(\mathbf{X}|\theta)$ мы оптимизируем ELBO $\mathcal{L}(q, \theta)$ как по θ, так и по $q(\mathbf{Z})$
- Алгоритм блочно-координатного спуска известен как EM-алгоритм

- Мы получаем

$$\log p(\mathbf{X}|\boldsymbol{\theta}) =$$

$$= \underbrace{\int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z}}_{\text{Нижняя оценка обусловленности ELBO } \mathcal{L}(q, \boldsymbol{\theta})} + \underbrace{\int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} d\mathbf{Z}}_{\text{Неотрицательное}}$$

- Таким образом

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p) \geq \mathcal{L}(q, \boldsymbol{\theta})$$

- Вместо оптимизации $\log p(\mathbf{X}|\boldsymbol{\theta})$ мы оптимизируем ELBO $\mathcal{L}(q, \boldsymbol{\theta})$ как по $\boldsymbol{\theta}$, так и по $q(\mathbf{Z})$
- Алгоритм блочно-координатного спуска известен как EM-алгоритм

- Мы получаем

$$\log p(\mathbf{X}|\boldsymbol{\theta}) =$$

$$= \underbrace{\int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z}}_{\text{Нижняя оценка обусловленности ELBO } \mathcal{L}(q, \boldsymbol{\theta})} + \underbrace{\int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} d\mathbf{Z}}_{\text{Неотрицательное}}$$

- Таким образом

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p) \geq \mathcal{L}(q, \boldsymbol{\theta})$$

- Вместо оптимизации $\log p(\mathbf{X}|\boldsymbol{\theta})$ мы оптимизируем ELBO $\mathcal{L}(q, \boldsymbol{\theta})$ как по $\boldsymbol{\theta}$, так и по $q(\mathbf{Z})$
- Алгоритм блочно-координатного спуска известен как EM-алгоритм

Вариационная нижняя граница

- Функция $g(\xi, \mathbf{x})$ называется вариационной нижней границей для $f(\mathbf{x})$ тогда и только тогда, когда
  - Для всех ξ и для всех $\mathbf{x}$ выполняется $f(\mathbf{x}) \geq g(\xi, \mathbf{x})$
  - Для любого $\mathbf{x}_0$ существует $\xi(\mathbf{x}_0)$ такое, что $f(\mathbf{x}_0) = g(\xi(\mathbf{x}_0), \mathbf{x}_0)$
- Если нам удалось найти вариационную нижнюю границу, то вместо решения

$$f(\mathbf{x}) \to \max_{\mathbf{x}}$$

мы можем итеративно вычислять координаты $g(\xi, \mathbf{x})$

$$\mathbf{x}_i = \arg\max_{\mathbf{x}} g(\xi_{i-1}, \mathbf{x}), \;\; \xi_i = \xi(\mathbf{x}_i) = \arg\max_{\xi} g(\xi, \mathbf{x}_i)$$

Вариационная нижняя граница

- Функция $g(\xi, \mathbf{x})$ называется вариационной нижней границей для $f(\mathbf{x})$ тогда и только тогда, когда
  - Для всех ξ и для всех $\mathbf{x}$ выполняется $f(\mathbf{x}) \geq g(\xi, \mathbf{x})$
  - Для любого $\mathbf{x}_0$ существует $\xi(\mathbf{x}_0)$ такое, что $f(\mathbf{x}_0) = g(\xi(\mathbf{x}_0), \mathbf{x}_0)$
- Если нам удалось найти вариационную нижнюю границу, то вместо решения

$$f(\mathbf{x}) \to \max_{\mathbf{x}}$$

мы можем итеративно вычислять координаты $g(\xi, \mathbf{x})$

$$\mathbf{x}_i = \arg\max_{\mathbf{x}} g(\xi_{i-1}, \mathbf{x}), \;\; \xi_i = \xi(\mathbf{x}_i) = \arg\max_{\xi} g(\xi, \mathbf{x}_i)$$

- Решение

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \to \max_{q, \boldsymbol{\theta}}$$

начинается с $\boldsymbol{\theta}_0$ и итеративно повторяется оптимизация по q и $\boldsymbol{\theta}$

- Найдем $\arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0)$.

- Мы получаем, что

$$\arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0) = \arg\min KL(q\|p) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0)$$

- Решение

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \to \max_{q, \boldsymbol{\theta}}$$

  начинается с $\boldsymbol{\theta}_0$ и итеративно повторяется оптимизация по q и $\boldsymbol{\theta}$
- Найдем $\arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0)$.

- Мы получаем, что

$$\arg\max_q \mathcal{L}(q, \theta_0) = \arg\min KL(q\|p) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0)$$

- Решение

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \to \max_{q, \boldsymbol{\theta}}$$

  начинается с $\boldsymbol{\theta}_0$ и итеративно повторяется оптимизация по q и $\boldsymbol{\theta}$
- Найдем $\arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0)$. Поскольку $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \cdot p(\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})$
- Мы получаем, что

$$\arg\max_q \mathcal{L}(q, \theta_0) = \arg\min KL(q\|p) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0)$$

- Решение

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \to \max_{q, \boldsymbol{\theta}}$$

начинается с $\boldsymbol{\theta}_0$ и итеративно повторяется оптимизация по q и $\boldsymbol{\theta}$

- Найдем $\arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0)$. Поскольку $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \cdot p(\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})$

$$\mathcal{L}(q, \boldsymbol{\theta}_0) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}_0)}{q(\mathbf{Z})} d\mathbf{Z}$$

- Мы получаем, что

$$\arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0) = \arg\min KL(q\|p) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0)$$

- Решение

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \to \max_{q, \boldsymbol{\theta}}$$

начинается с $\boldsymbol{\theta}_0$ и итеративно повторяется оптимизация по q и $\boldsymbol{\theta}$

- Найдем $\arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0)$. Поскольку $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \cdot p(\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})$

$$\mathcal{L}(q, \boldsymbol{\theta}_0) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}_0)}{q(\mathbf{Z})} d\mathbf{Z}$$

$$= \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0) \cdot p(\mathbf{X}|\boldsymbol{\theta}_0)}{q(\mathbf{Z})} d\mathbf{Z}$$

- Мы получаем, что

$$\arg\max_q \mathcal{L}(q, \theta_0) = \arg\min KL(q||p) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0)$$

- Решение

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \to \max_{q, \boldsymbol{\theta}}$$

начинается с $\boldsymbol{\theta}_0$ и итеративно повторяется оптимизация по q и $\boldsymbol{\theta}$

- Найдем $\arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0)$. Поскольку $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \cdot p(\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})$

$$\begin{aligned}
\mathcal{L}(q, \boldsymbol{\theta}_0) &= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}_0)}{q(\mathbf{Z})} d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0) \cdot p(\mathbf{X}|\boldsymbol{\theta}_0)}{q(\mathbf{Z})} d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}_0) d\mathbf{Z}
\end{aligned}$$

- Мы получаем, что

$$\arg\max_q \mathcal{L}(q, \theta_0) = \arg\min KL(q||p) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0)$$

- Решение

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \to \max_{q, \boldsymbol{\theta}}$$

начинается с $\boldsymbol{\theta}_0$ и итеративно повторяется оптимизация по q и $\boldsymbol{\theta}$

- Найдем $\arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0)$. Поскольку $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \cdot p(\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})$

$$\begin{aligned}
\mathcal{L}(q, \boldsymbol{\theta}_0) &= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}_0)}{q(\mathbf{Z})} d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0) \cdot p(\mathbf{X}|\boldsymbol{\theta}_0)}{q(\mathbf{Z})} d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}_0) d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} + \log p(\mathbf{X}|\boldsymbol{\theta}_0)
\end{aligned}$$

- Мы получаем, что

$$\arg\max_q \mathcal{L}(q, \theta_0) = \arg\min KL(q||p) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0)$$

- Решение

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \to \max_{q, \boldsymbol{\theta}}$$

начинается с $\boldsymbol{\theta}_0$ и итеративно повторяется оптимизация по q и $\boldsymbol{\theta}$

- Найдем $\arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0)$. Поскольку $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \cdot p(\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})$

$$\begin{aligned}
\mathcal{L}(q, \boldsymbol{\theta}_0) &= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}_0)}{q(\mathbf{Z})} d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0) \cdot p(\mathbf{X}|\boldsymbol{\theta}_0)}{q(\mathbf{Z})} d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}_0) d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} + \log p(\mathbf{X}|\boldsymbol{\theta}_0) \\
&= -KL(q\|p) + \log p(\mathbf{X}|\boldsymbol{\theta}_0)
\end{aligned}$$

- Мы получаем, что

$$\arg\max_q \mathcal{L}(q, \theta_0) = \arg\min KL(q\|p) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0)$$

- Решение

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \to \max_{q, \boldsymbol{\theta}}$$

начинается с $\boldsymbol{\theta}_0$ и итеративно повторяется оптимизация по q и $\boldsymbol{\theta}$

- Найдем $\arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0)$. Поскольку $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \cdot p(\mathbf{X}|\boldsymbol{\theta}) = p(\mathbf{Z}, \mathbf{X}|\boldsymbol{\theta})$

$$\begin{aligned}
\mathcal{L}(q, \boldsymbol{\theta}_0) &= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}_0)}{q(\mathbf{Z})} d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0) \cdot p(\mathbf{X}|\boldsymbol{\theta}_0)}{q(\mathbf{Z})} d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}_0) d\mathbf{Z} \\
&= \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} + \log p(\mathbf{X}|\boldsymbol{\theta}_0) \\
&= -KL(q\|p) + \log p(\mathbf{X}|\boldsymbol{\theta}_0)
\end{aligned}$$

- Мы получаем, что

$$\arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0) = \arg\min KL(q\|p) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0)$$

- Таким образом, чтобы решить

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \to \max_{q, \boldsymbol{\theta}}$$

мы начинаем с начального $\boldsymbol{\theta}_0$ и повторяем итеративно

- **E-шаг**: найти

$$q(\mathbf{Z}) = \arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0) = \arg\min_q KL(q||p) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0)$$

- **M-шаг**: решить

$$\boldsymbol{\theta}_* = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}),$$

установить $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_*$ и вернуться к **E-шагу** до сходимости алгоритма

- EM-алгоритм монотонно увеличивает нижнюю границу и сходится к стационарной точке $\log p(\mathbf{X}|\boldsymbol{\theta})$

- Таким образом, чтобы решить

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \to \max_{q, \boldsymbol{\theta}}$$

мы начинаем с начального $\boldsymbol{\theta}_0$ и повторяем итеративно

- **E-шаг**: найти

$$q(\mathbf{Z}) = \arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0) = \arg\min_q KL(q||p) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0)$$

- **M-шаг**: решить

$$\boldsymbol{\theta}_* = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}),$$

установить $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_*$ и вернуться к **E-шагу** до сходимости алгоритма

- EM-алгоритм монотонно увеличивает нижнюю границу и сходится к стационарной точке $\log p(\mathbf{X}|\boldsymbol{\theta})$

- Таким образом, чтобы решить

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \to \max_{q, \boldsymbol{\theta}}$$

мы начинаем с начального $\boldsymbol{\theta}_0$ и повторяем итеративно

- **E-шаг**: найти

$$q(\mathbf{Z}) = \arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0) = \arg\min_q KL(q||p) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0)$$

- **M-шаг**: решить

$$\boldsymbol{\theta}_* = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}),$$

установить $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_*$ и вернуться к **E-шагу** до сходимости алгоритма

- EM-алгоритм монотонно увеличивает нижнюю границу и сходится к стационарной точке $\log p(\mathbf{X}|\boldsymbol{\theta})$

- Таким образом, чтобы решить

$$\mathcal{L}(q, \boldsymbol{\theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \to \max_{q, \boldsymbol{\theta}}$$

мы начинаем с начального $\boldsymbol{\theta}_0$ и повторяем итеративно

- **E-шаг**: найти

$$q(\mathbf{Z}) = \arg\max_q \mathcal{L}(q, \boldsymbol{\theta}_0) = \arg\min_q KL(q||p) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_0)$$

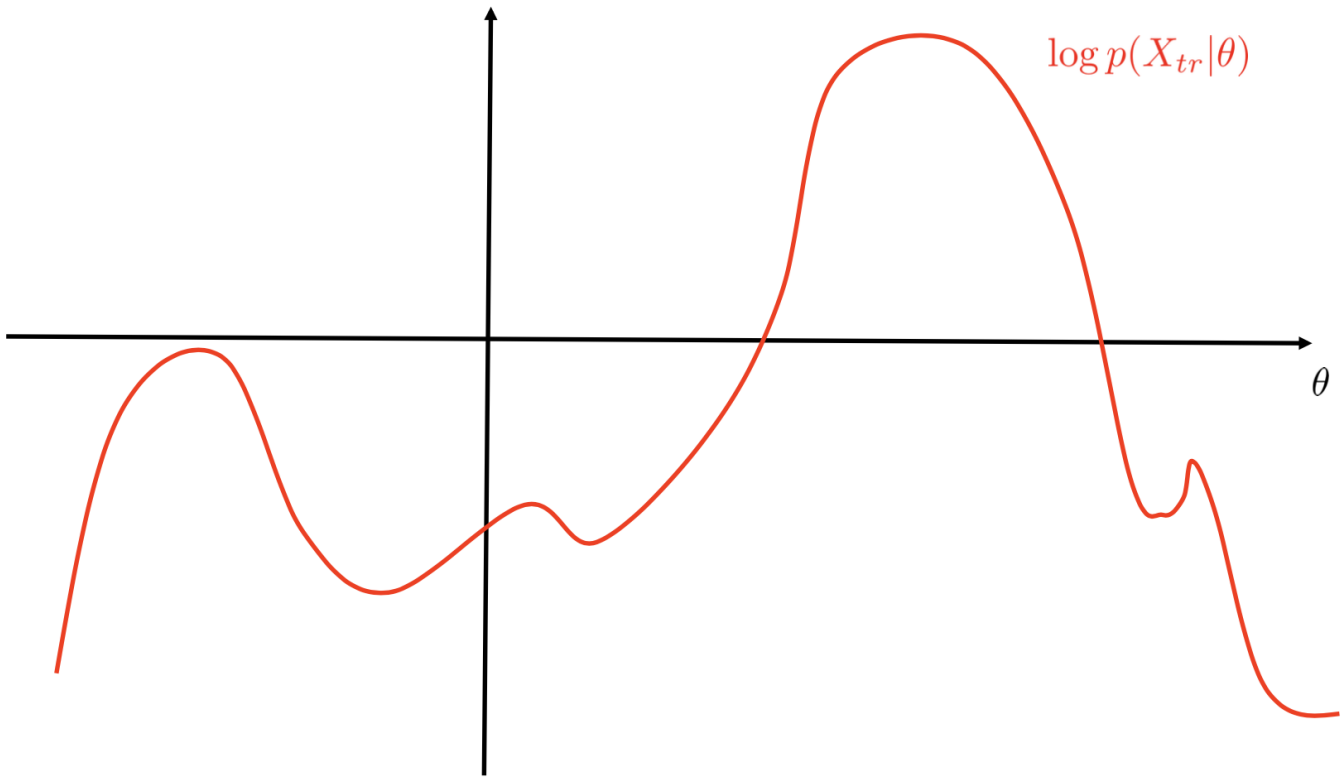
- **M-шаг**: решить

$$\boldsymbol{\theta}_* = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}),$$

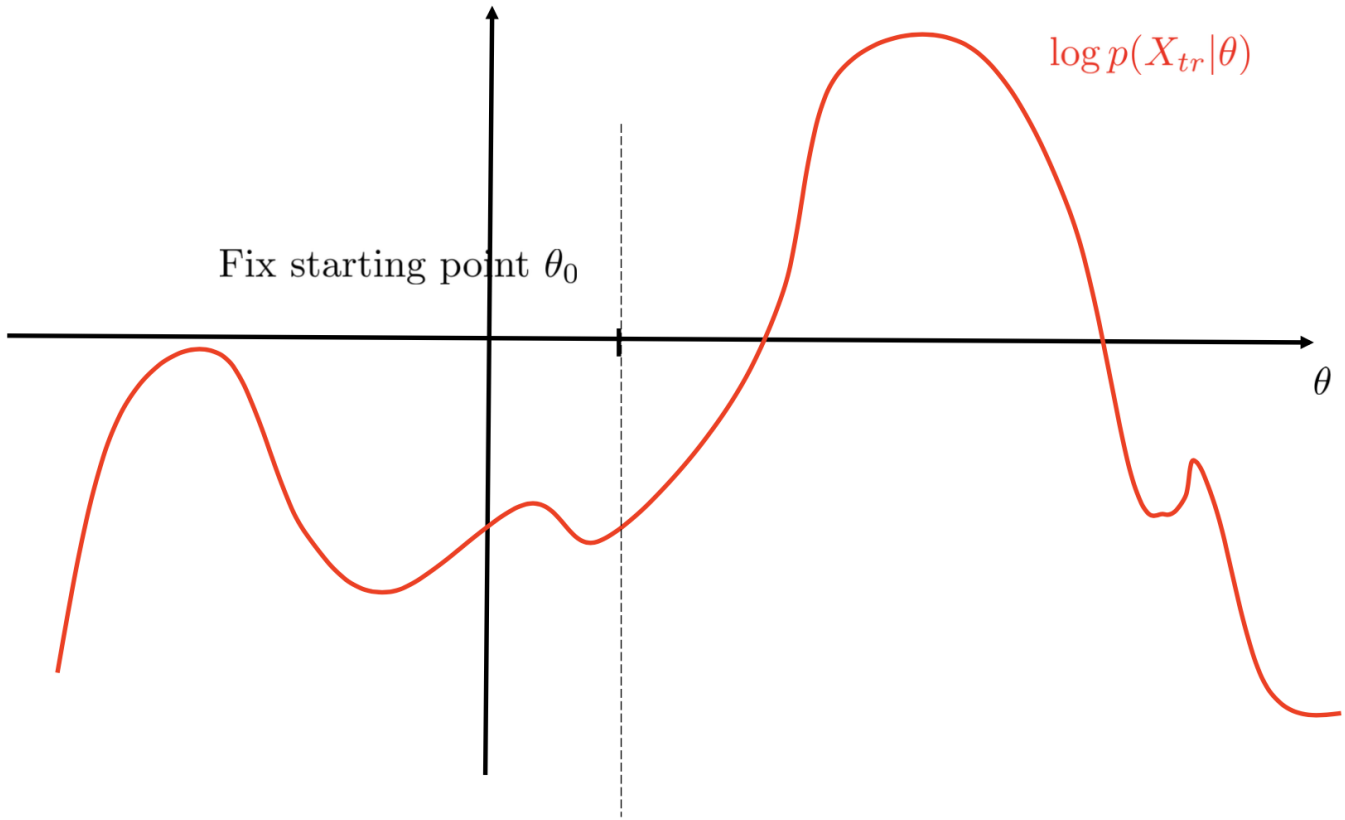
установить $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_*$ и вернуться к **E-шагу** до сходимости алгоритма

- EM-алгоритм монотонно увеличивает нижнюю границу и сходится к стационарной точке $\log p(\mathbf{X}|\boldsymbol{\theta})$

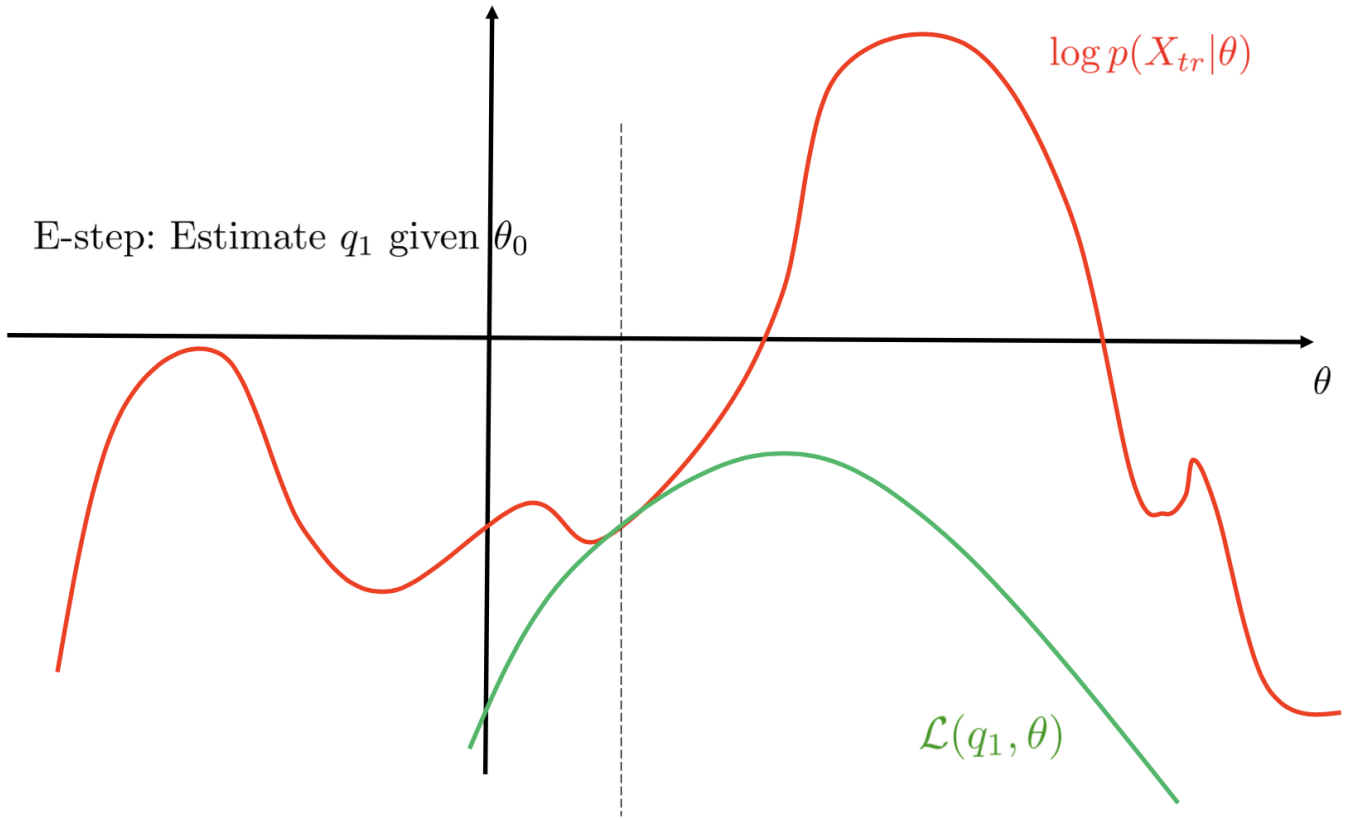
EM-алгоритм



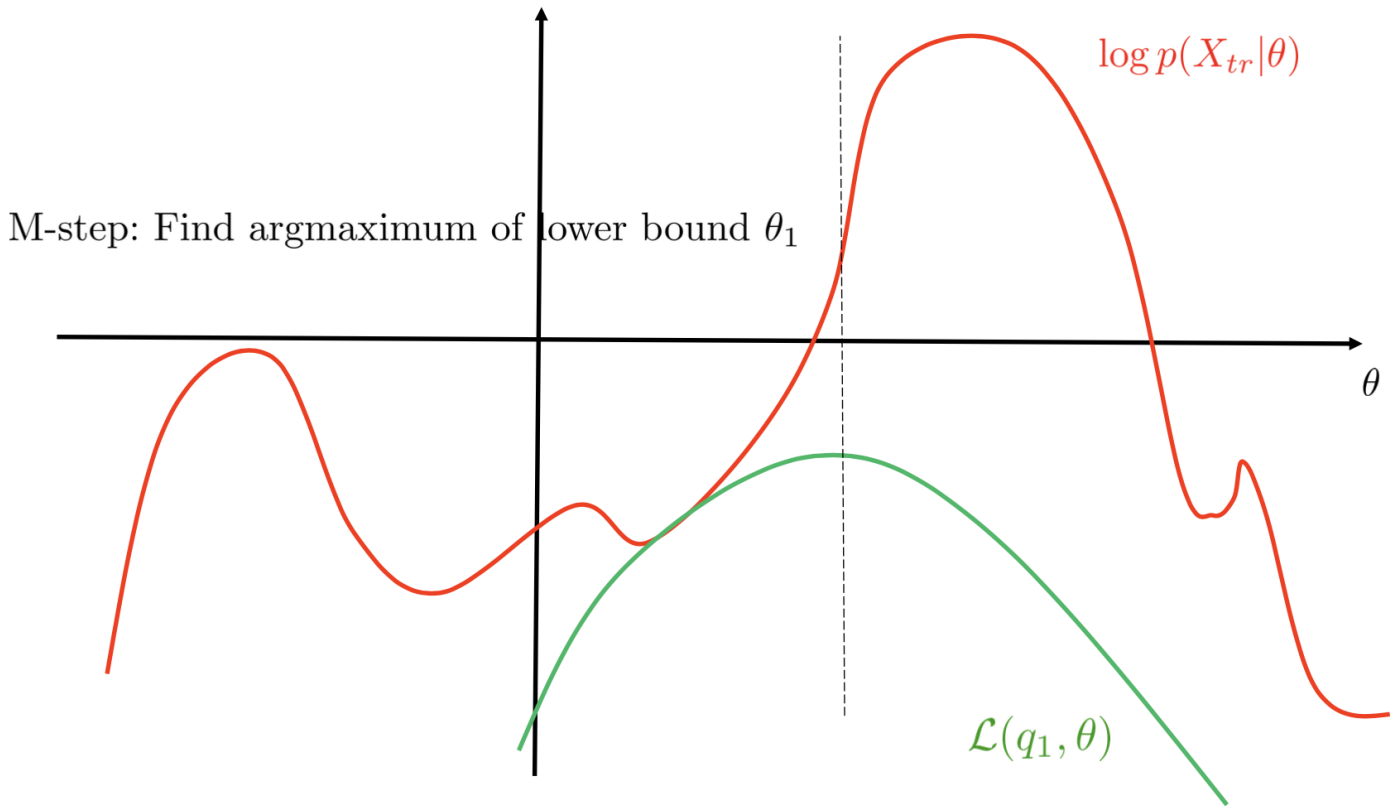
EM-алгоритм



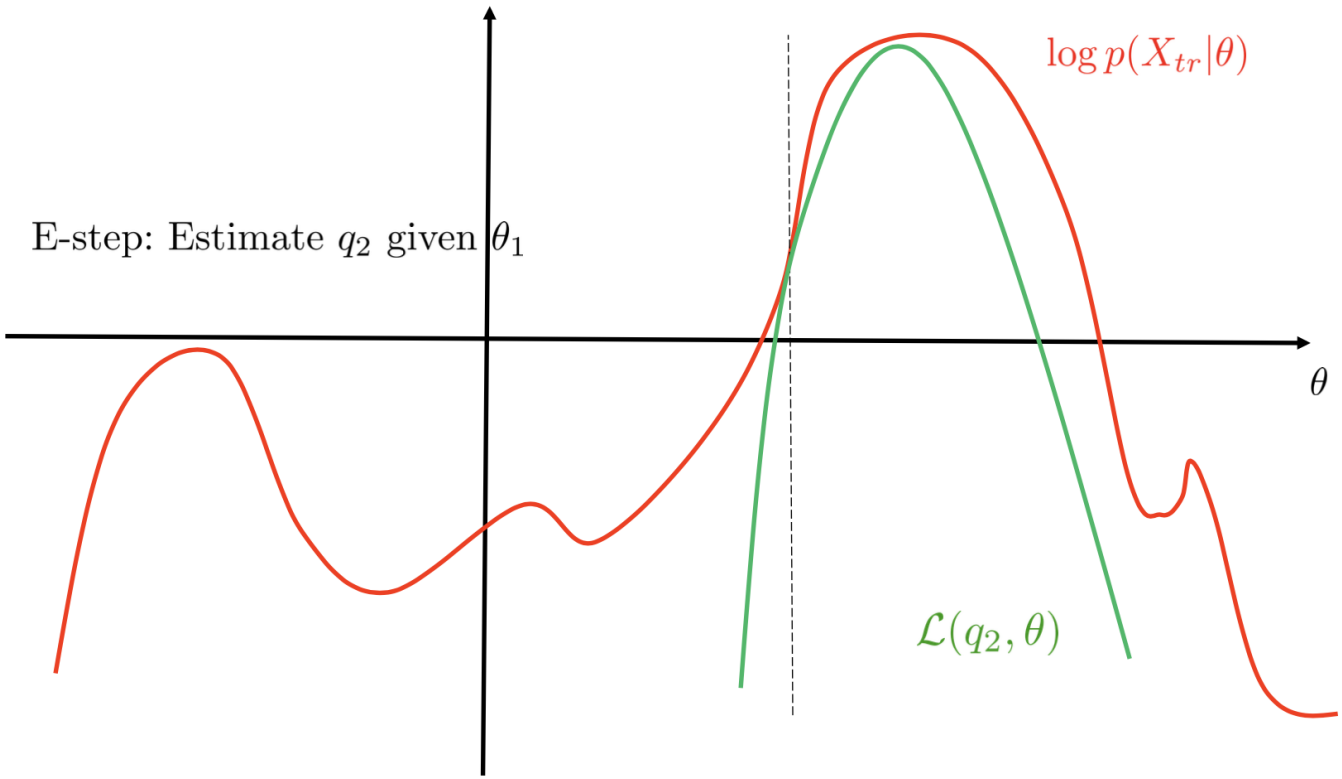
EM-алгоритм


$\log p(X_{tr}|\theta)$
E-step: Estimate q_1 given θ_0
θ
$\mathcal{L}(q_1, \theta)$

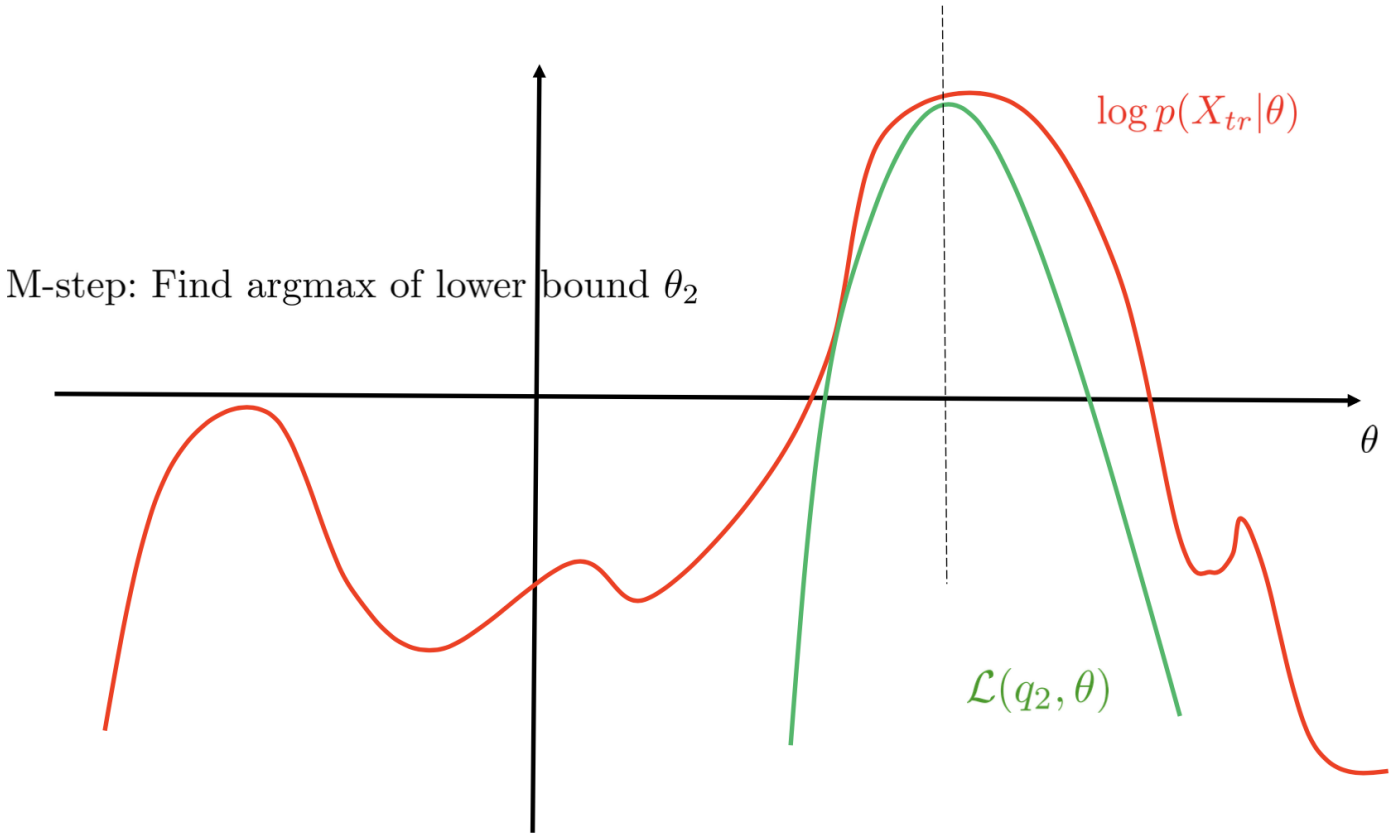
EM-алгоритм


M-step: Find argmaximum of lower bound θ_1
$\log p(X_{tr}|\theta)$
θ
$\mathcal{L}(q_1, \theta)$

EM-алгоритм


E-step: Estimate q_2 given θ_1
$\log p(X_{tr}|\theta)$
$\mathcal{L}(q_2, \theta)$
θ

EM-алгоритм


M-step: Find argmax of lower bound θ_2
$\log p(X_{tr}|\theta)$
$\mathcal{L}(q_2, \theta)$
θ

Стоит ли использовать EM-алгоритм?

- Во многих случаях (например, для Гауссовской смеси) E-шаг и M-шаг могут выполняться в явном виде
- Позволяет строить более сложные модели данных, используя смеси простых распределений
- Если истинное апостериорное распределение $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ трудноразрешимо, мы можем искать ближайшее $q(\mathbf{Z})$ среди более «удобных» распределений путем решения задачи оптимизации
- Позволяет обрабатывать пропущенные данные, рассматривая их как скрытые переменные

Стоит ли использовать EM-алгоритм?

- Во многих случаях (например, для Гауссовской смеси) E-шаг и M-шаг могут выполняться в явном виде
- Позволяет строить более сложные модели данных, используя смеси простых распределений
- Если истинное апостериорное распределение $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ трудноразрешимо, мы можем искать ближайшее $q(\mathbf{Z})$ среди более «удобных» распределений путем решения задачи оптимизации
- Позволяет обрабатывать пропущенные данные, рассматривая их как скрытые переменные

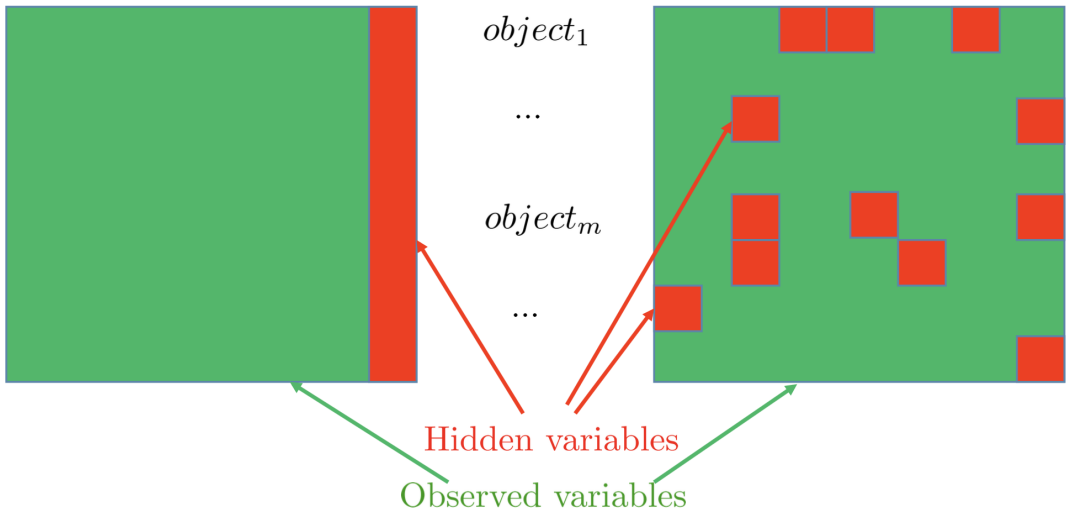
Стоит ли использовать EM-алгоритм?

- Во многих случаях (например, для Гауссовской смеси) E-шаг и M-шаг могут выполняться в явном виде
- Позволяет строить более сложные модели данных, используя смеси простых распределений
- Если истинное апостериорное распределение $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ трудноразрешимо, мы можем искать ближайшее $q(\mathbf{Z})$ среди более «удобных» распределений путем решения задачи оптимизации
- Позволяет обрабатывать пропущенные данные, рассматривая их как скрытые переменные

Стоит ли использовать EM-алгоритм?

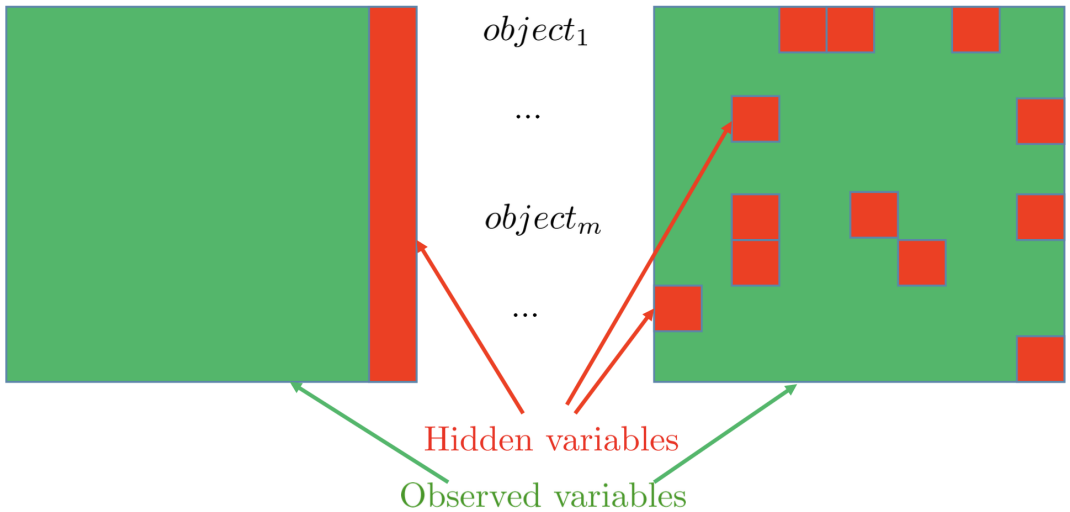
- Во многих случаях (например, для Гауссовской смеси) E-шаг и M-шаг могут выполняться в явном виде
- Позволяет строить более сложные модели данных, используя смеси простых распределений
- Если истинное апостериорное распределение $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ трудноразрешимо, мы можем искать ближайшее $q(\mathbf{Z})$ среди более «удобных» распределений путем решения задачи оптимизации
- Позволяет обрабатывать пропущенные данные, рассматривая их как скрытые переменные

Стоит ли использовать EM?



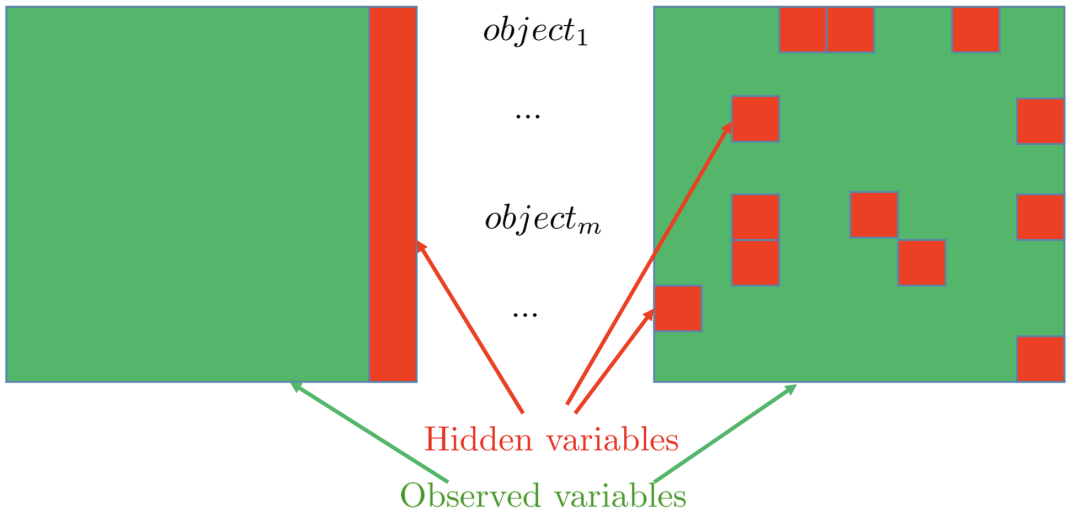
- EM-алгоритм позволяет заполнить произвольные пропуски в данных
- Может работать как с дискретными, так и с непрерывными переменными
- Всегда сходится
- Позволяет работать в многомерном пространстве

Стоит ли использовать EM?



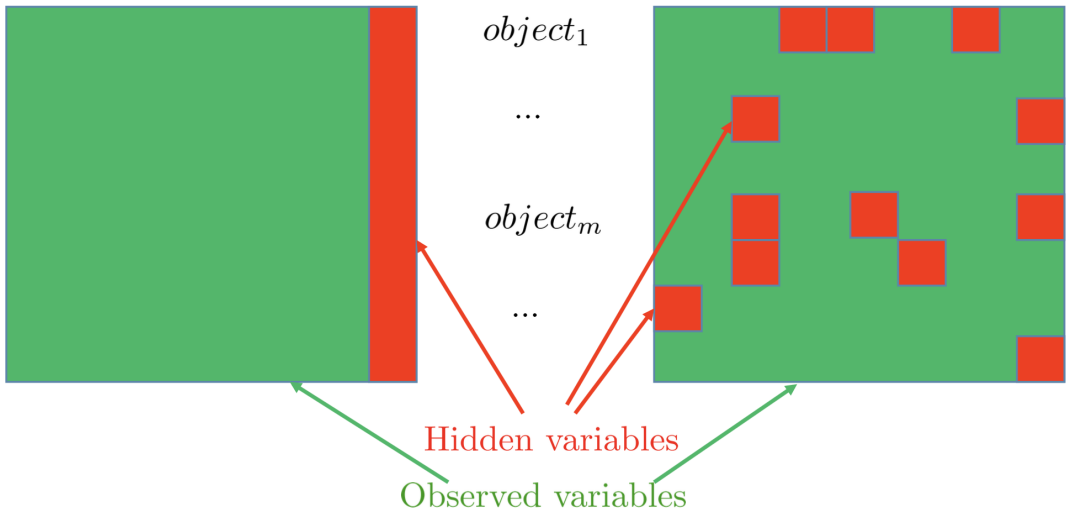
- EM-алгоритм позволяет заполнить произвольные пропуски в данных
- Может работать как с дискретными, так и с непрерывными переменными
- Всегда сходится
- Позволяет работать в многомерном пространстве

Стоит ли использовать EM?



- EM-алгоритм позволяет заполнить произвольные пропуски в данных
- Может работать как с дискретными, так и с непрерывными переменными
- Всегда сходится
- Позволяет работать в многомерном пространстве

Стоит ли использовать EM?



- EM-алгоритм позволяет заполнить произвольные пропуски в данных
- Может работать как с дискретными, так и с непрерывными переменными
- Всегда сходится
- Позволяет работать в многомерном пространстве

Категориальные скрытые переменные

- Предположим, что все $z_i \in \{1, \ldots, K\}$ тогда маргинальное распределение

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} p(\mathbf{x}_i|k, \boldsymbol{\theta}) p(z_i = k|\boldsymbol{\theta})$$

  является конечной смесью распределений

- E-шаг может быть выписан в явном виде

$$q(z_i = k) = p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_i|k, \boldsymbol{\theta}) p(z_i = k|\boldsymbol{\theta})}{\sum_{l=1}^{K} p(\mathbf{x}_i|l, \boldsymbol{\theta}) p(z_i = l|\boldsymbol{\theta})}$$

- M-шаг - это просто сумма конечных членов

$$\mathbb{E}_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \sum_{i=1}^{m} \mathbb{E}_{z_i} \log p(x_i, z_i|\boldsymbol{\theta}) =$$

$$\sum_{i=1}^{m} \sum_{k=1}^{K} q(z_i = k) \log p(x_i, k|\boldsymbol{\theta})$$

Категориальные скрытые переменные

- Предположим, что все $z_i \in \{1, \dots, K\}$ тогда маргинальное распределение

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} p(\mathbf{x}_i|k, \boldsymbol{\theta}) p(z_i = k|\boldsymbol{\theta})$$

  является конечной смесью распределений
- E-шаг может быть выписан в явном виде

$$q(z_i = k) = p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_i|k, \boldsymbol{\theta}) p(z_i = k|\boldsymbol{\theta})}{\sum_{l=1}^{K} p(\mathbf{x}_i|l, \boldsymbol{\theta}) p(z_i = l|\boldsymbol{\theta})}$$

- M-шаг - это просто сумма конечных членов

$$\mathbb{E}_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \sum_{i=1}^{m} \mathbb{E}_{z_i} \log p(x_i, z_i|\boldsymbol{\theta}) =$$

$$\sum_{i=1}^{m} \sum_{k=1}^{K} q(z_i = k) \log p(x_i, k|\boldsymbol{\theta})$$

Категориальные скрытые переменные

- Предположим, что все $z_i \in \{1, \dots, K\}$ тогда маргинальное распределение

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} p(\mathbf{x}_i|k, \boldsymbol{\theta}) p(z_i = k|\boldsymbol{\theta})$$

  является конечной смесью распределений
- E-шаг может быть выписан в явном виде

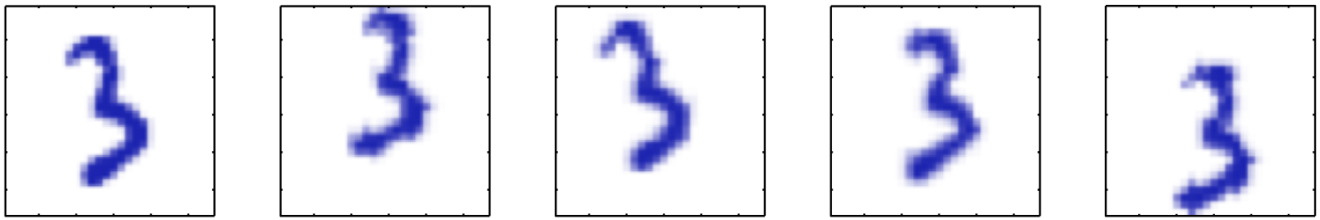
$$q(z_i = k) = p(z_i = k|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_i|k, \boldsymbol{\theta}) p(z_i = k|\boldsymbol{\theta})}{\sum_{l=1}^{K} p(\mathbf{x}_i|l, \boldsymbol{\theta}) p(z_i = l|\boldsymbol{\theta})}$$

- M-шаг - это просто сумма конечных членов

$$\mathbb{E}_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \sum_{i=1}^{m} \mathbb{E}_{z_i} \log p(x_i, z_i|\boldsymbol{\theta}) =$$

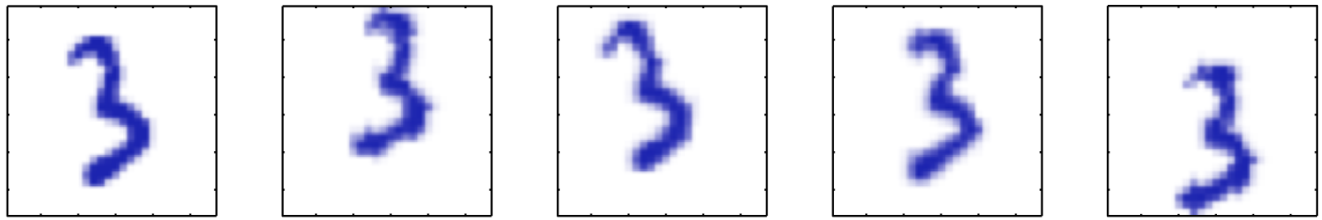
$$\sum_{i=1}^{m} \sum_{k=1}^{K} q(z_i = k) \log p(x_i, k|\boldsymbol{\theta})$$

Многообразия



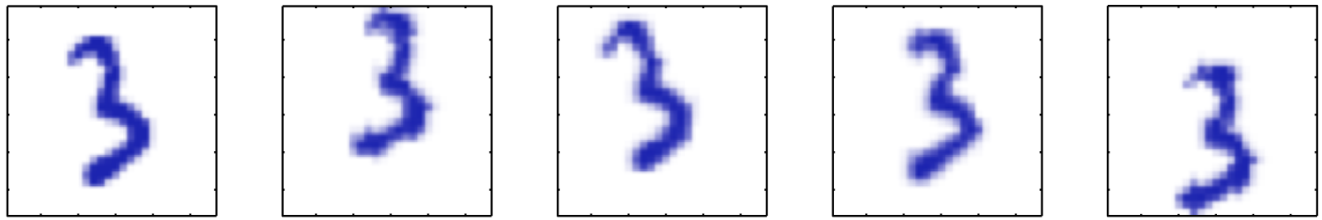
- Реальные наборы данных: точки данных лежат близко к многообразию гораздо меньшей размерности
- Изображение 100×100 в сером цвете т.е. 10^4 - мерное пространство данных
- три степени изменчивости: вертикальные / горизонтальные перемещения и вращения, описанные некоторыми скрытыми переменными
- трехмерное нелинейное многообразие
- данные реальных цифр: дополнительные степени свободы от масштабирования, изменчивости индивидуального письма, стилей письма
- На практике точки данных не будут сильно ограничены гладким низкоразмерным многообразием: могут интерпретироваться как шум

Многообразия



- Реальные наборы данных: точки данных лежат близко к многообразию гораздо меньшей размерности
- Изображение 100×100 в сером цвете т.е. 10^4 - мерное пространство данных
- три степени изменчивости: вертикальные / горизонтальные перемещения и вращения, описанные некоторыми скрытыми переменными
- трехмерное нелинейное многообразие
- данные реальных цифр: дополнительные степени свободы от масштабирования, изменчивости индивидуального письма, стилей письма
- На практике точки данных не будут сильно ограничены гладким низкоразмерным многообразием: могут интерпретироваться как шум

Многообразия



- Реальные наборы данных: точки данных лежат близко к многообразию гораздо меньшей размерности
- Изображение 100×100 в сером цвете т.е. 10^4 - мерное пространство данных
- три степени изменчивости: вертикальные / горизонтальные перемещения и вращения, описанные некоторыми скрытыми переменными
- трехмерное нелинейное многообразие
- данные реальных цифр: дополнительные степени свободы от масштабирования, изменчивости индивидуального письма, стилей письма
- На практике точки данных не будут сильно ограничены гладким низкоразмерным многообразием: могут интерпретироваться как шум

Непрерывные скрытые переменные

- Непрерывные переменные можно рассматривать как бесконечную смесь распределений

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \int p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta})d\mathbf{z}_i = \int p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta})p(\mathbf{z}_i|\boldsymbol{\theta})d\mathbf{z}_i$$

- E-шаг может быть выписан в явном виде только в случае сопряженных распределений

$$q(\mathbf{z}_i) = p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta})p(\mathbf{z}_i|\boldsymbol{\theta})}{\int p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta})p(\mathbf{z}_i|\boldsymbol{\theta})d\mathbf{z}_i}$$

- Обычно непрерывные скрытые переменные используются для уменьшения размерности, также известной как обучение представлению

Непрерывные скрытые переменные

- Непрерывные переменные можно рассматривать как бесконечную смесь распределений

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \int p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta})d\mathbf{z}_i = \int p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta})p(\mathbf{z}_i|\boldsymbol{\theta})d\mathbf{z}_i$$

- E-шаг может быть выписан в явном виде только в случае сопряженных распределений

$$q(\mathbf{z}_i) = p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta})p(\mathbf{z}_i|\boldsymbol{\theta})}{\int p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta})p(\mathbf{z}_i|\boldsymbol{\theta})d\mathbf{z}_i}$$

- Обычно непрерывные скрытые переменные используются для уменьшения размерности, также известной как обучение представлению

Непрерывные скрытые переменные

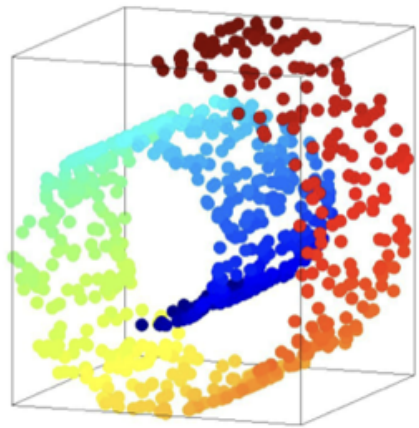
- Непрерывные переменные можно рассматривать как бесконечную смесь распределений

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \int p(\mathbf{x}_i, \mathbf{z}_i|\boldsymbol{\theta})d\mathbf{z}_i = \int p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta})p(\mathbf{z}_i|\boldsymbol{\theta})d\mathbf{z}_i$$

- E-шаг может быть выписан в явном виде только в случае сопряженных распределений

$$q(\mathbf{z}_i) = p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta})p(\mathbf{z}_i|\boldsymbol{\theta})}{\int p(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\theta})p(\mathbf{z}_i|\boldsymbol{\theta})d\mathbf{z}_i}$$

- Обычно непрерывные скрытые переменные используются для уменьшения размерности, также известной как обучение представлению



Методология EM

- Разработка вероятностной параметрической модели данных
- Включение дополнительных (скрытых) переменных, пока модель не станет достаточно простой, например, принадлежащей экспоненциальному классу
- Обработка всех пропущенных значений в данных как скрытых переменных
- Использование EM при подгонке модели к данным (например, с использованием MLE)
- Оценка распределения по скрытым переменным
- Максимизация ожидания в соответствии со скрытым переменным логарифма совместного правдоподобия по параметрам

Методология EM

- Разработка вероятностной параметрической модели данных
- Включение дополнительных (скрытых) переменных, пока модель не станет достаточно простой, например, принадлежащей экспоненциальному классу
- Обработка всех пропущенных значений в данных как скрытых переменных
- Использование EM при подгонке модели к данным (например, с использованием MLE)
- Оценка распределения по скрытым переменным
- Максимизация ожидания в соответствии со скрытым переменным логарифма совместного правдоподобия по параметрам

Методология EM

- Разработка вероятностной параметрической модели данных
- Включение дополнительных (скрытых) переменных, пока модель не станет достаточно простой, например, принадлежащей экспоненциальному классу
- Обработка всех пропущенных значений в данных как скрытых переменных
- Использование EM при подгонке модели к данным (например, с использованием MLE)
- Оценка распределения по скрытым переменным
- Максимизация ожидания в соответствии со скрытым переменным логарифма совместного правдоподобия по параметрам

Методология EM

- Разработка вероятностной параметрической модели данных
- Включение дополнительных (скрытых) переменных, пока модель не станет достаточно простой, например, принадлежащей экспоненциальному классу
- Обработка всех пропущенных значений в данных как скрытых переменных
- Использование EM при подгонке модели к данным (например, с использованием MLE)
- Оценка распределения по скрытым переменным
- Максимизация ожидания в соответствии со скрытым переменным логарифма совместного правдоподобия по параметрам

Методология EM

- Разработка вероятностной параметрической модели данных
- Включение дополнительных (скрытых) переменных, пока модель не станет достаточно простой, например, принадлежащей экспоненциальному классу
- Обработка всех пропущенных значений в данных как скрытых переменных
- Использование EM при подгонке модели к данным (например, с использованием MLE)
- Оценка распределения по скрытым переменным
- Максимизация ожидания в соответствии со скрытым переменным логарифма совместного правдоподобия по параметрам

Методология EM

- Разработка вероятностной параметрической модели данных
- Включение дополнительных (скрытых) переменных, пока модель не станет достаточно простой, например, принадлежащей экспоненциальному классу
- Обработка всех пропущенных значений в данных как скрытых переменных
- Использование EM при подгонке модели к данным (например, с использованием MLE)
- Оценка распределения по скрытым переменным
- Максимизация ожидания в соответствии со скрытым переменным логарифма совместного правдоподобия по параметрам

Проблемы с EM

- Каждый объект имеет многомерную дискретную скрытую переменную $\Rightarrow$ экспоненциально большие суммы
- Объект имеет как дискретные, так и непрерывные скрытые переменные (например, смесь низкоразмерных многообразий) $\Rightarrow$ смешанные дискретно-непрерывные распределения по скрытым переменным
- Непрерывные латентные переменные происходят из несопряженных априорных распределений $\Rightarrow$ невычислимые многомерные интегралы
- Дальнейший подход: крупномасштабный вариационный байесовский метод

Проблемы с EM

- Каждый объект имеет многомерную дискретную скрытую переменную $\Rightarrow$ экспоненциально большие суммы
- Объект имеет как дискретные, так и непрерывные скрытые переменные (например, смесь низкоразмерных многообразий) $\Rightarrow$ смешанные дискретно-непрерывные распределения по скрытым переменным
- Непрерывные латентные переменные происходят из несопряженных априорных распределений $\Rightarrow$ невычислимые многомерные интегралы
- Дальнейший подход: крупномасштабный вариационный байесовский метод

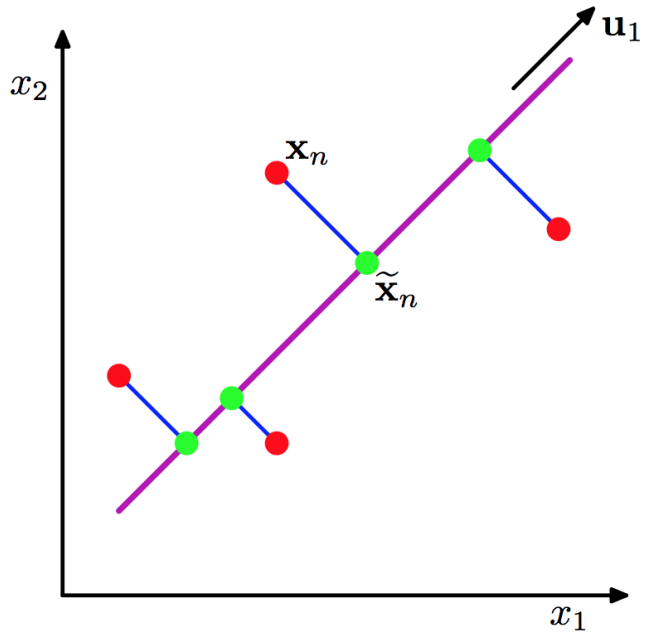
Проблемы с EM

- Каждый объект имеет многомерную дискретную скрытую переменную $\Rightarrow$ экспоненциально большие суммы
- Объект имеет как дискретные, так и непрерывные скрытые переменные (например, смесь низкоразмерных многообразий) $\Rightarrow$ смешанные дискретно-непрерывные распределения по скрытым переменным
- Непрерывные латентные переменные происходят из несопряженных априорных распределений $\Rightarrow$ невычислимые многомерные интегралы
- Дальнейший подход: крупномасштабный вариационный байесовский метод

Проблемы с EM

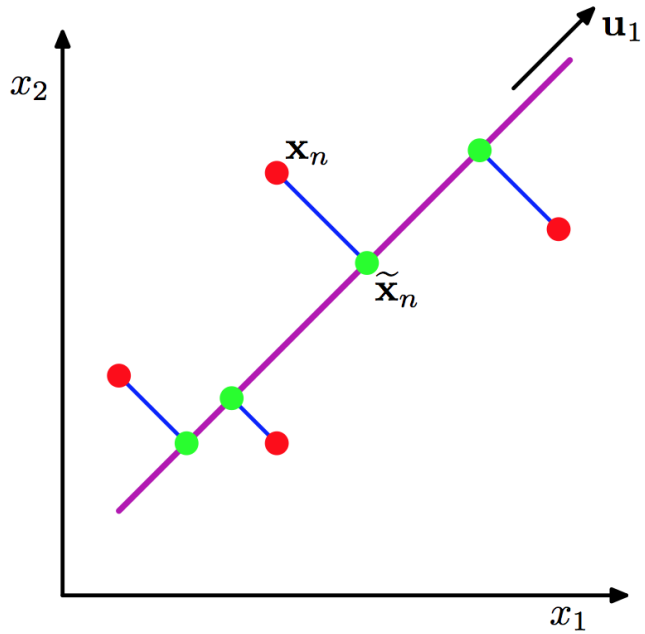
- Каждый объект имеет многомерную дискретную скрытую переменную $\Rightarrow$ экспоненциально большие суммы
- Объект имеет как дискретные, так и непрерывные скрытые переменные (например, смесь низкоразмерных многообразий) $\Rightarrow$ смешанные дискретно-непрерывные распределения по скрытым переменным
- Непрерывные латентные переменные происходят из несопряженных априорных распределений $\Rightarrow$ невычислимые многомерные интегралы
- Дальнейший подход: крупномасштабный вариационный байесовский метод

Формулировка подхода через максимальную дисперсию



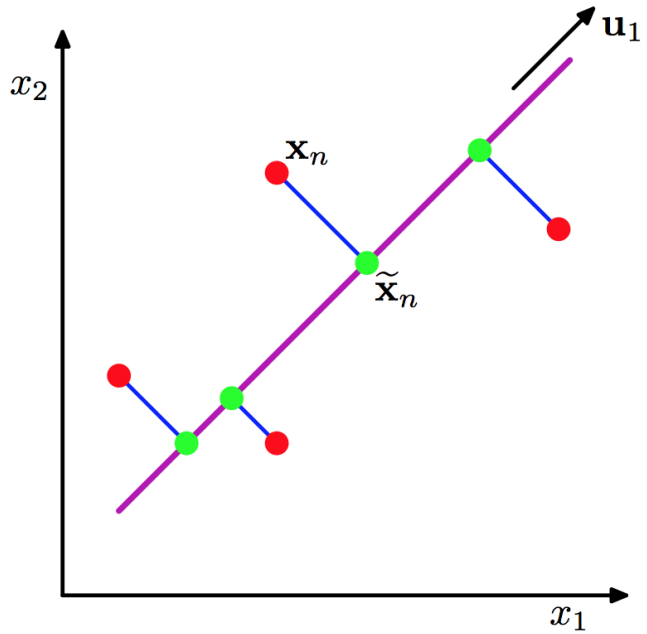
- $\{\mathbf{x}_i\}_{i=1}^m$, $\mathbf{x}_i \in \mathbb{R}^d$ образец данных
- Цель: спроецировать данные в пространство с размерностью $q < d$, в то же время максимизируя дисперсию проецируемых точек
- Пусть $q = 1$ и через $\mathbf{u}_1 \in \mathbb{R}^d$ обозначим d-мерный вектор, такой что $\mathbf{u}_1^\top \mathbf{u}_1 = 1$

Формулировка подхода через максимальную дисперсию



- $\{\mathbf{x}_i\}_{i=1}^m$, $\mathbf{x}_i \in \mathbb{R}^d$ образец данных
- Цель: спроецировать данные в пространство с размерностью $q < d$, в то же время максимизируя дисперсию проецируемых точек
- Пусть $q = 1$ и через $\mathbf{u}_1 \in \mathbb{R}^d$ обозначим d-мерный вектор, такой что $\mathbf{u}_1^\top \mathbf{u}_1 = 1$

Формулировка подхода через максимальную дисперсию



- $\{\mathbf{x}_i\}_{i=1}^m$, $\mathbf{x}_i \in \mathbb{R}^d$ образец данных
- Цель: спроецировать данные в пространство с размерностью $q < d$, в то же время максимизируя дисперсию проецируемых точек
- Пусть $q = 1$ и через $\mathbf{u}_1 \in \mathbb{R}^d$ обозначим d-мерный вектор, такой что $\mathbf{u}_1^\top \mathbf{u}_1 = 1$

Подход через максимальную дисперсию

- Если мы обозначим через $\overline{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$, то дисперсия проецируемых данных

$$\frac{1}{m}\sum_{i=1}^{m}\{\mathbf{u}_1^\top\mathbf{x}_i - \mathbf{u}_1^\top\overline{\mathbf{x}}\}^2 = \mathbf{u}_1^\top\mathbf{S}\mathbf{u}_1,$$

  где $\mathbf{S} = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i - \overline{x})(\mathbf{x}_i - \overline{x})^\top$

- Приравняем производную $\mathbf{u}_1^\top\mathbf{S}\mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^\top\mathbf{u}_1)$ к нулю, получаем, что

$$\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1$$

- По индукции: оптимальные линейные проекции с максимальной дисперсией определяются q собственными векторами $\mathbf{u}_1, \ldots, \mathbf{u}_q$ ковариационной матрицы $\mathbf{S}$, соответствующей q наибольшим собственным значениям $\lambda_1, \ldots, \lambda_q$

Подход через максимальную дисперсию

- Если мы обозначим через $\overline{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$, то дисперсия проецируемых данных

$$\frac{1}{m}\sum_{i=1}^{m}\{\mathbf{u}_1^\top\mathbf{x}_i - \mathbf{u}_1^\top\overline{\mathbf{x}}\}^2 = \mathbf{u}_1^\top\mathbf{S}\mathbf{u}_1,$$

  где $\mathbf{S} = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i - \overline{x})(\mathbf{x}_i - \overline{x})^\top$
- Приравняем производную $\mathbf{u}_1^\top\mathbf{S}\mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^\top\mathbf{u}_1)$ к нулю, получаем, что

$$\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1$$

- По индукции: оптимальные линейные проекции с максимальной дисперсией определяются q собственными векторами $\mathbf{u}_1, \ldots, \mathbf{u}_q$ ковариационной матрицы $\mathbf{S}$, соответствующей q наибольшим собственным значениям $\lambda_1, \ldots, \lambda_q$

Подход через максимальную дисперсию

- Если мы обозначим через $\overline{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i$, то дисперсия проецируемых данных

$$\frac{1}{m}\sum_{i=1}^{m}\{\mathbf{u}_1^\top\mathbf{x}_i - \mathbf{u}_1^\top\overline{\mathbf{x}}\}^2 = \mathbf{u}_1^\top\mathbf{S}\mathbf{u}_1,$$

  где $\mathbf{S} = \frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_i - \overline{x})(\mathbf{x}_i - \overline{x})^\top$
- Приравняем производную $\mathbf{u}_1^\top\mathbf{S}\mathbf{u}_1 + \lambda_1(1 - \mathbf{u}_1^\top\mathbf{u}_1)$ к нулю, получаем, что

$$\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1$$

- По индукции: оптимальные линейные проекции с максимальной дисперсией определяются q собственными векторами $\mathbf{u}_1, \ldots, \mathbf{u}_q$ ковариационной матрицы $\mathbf{S}$, соответствующей q наибольшим собственным значениям $\lambda_1, \ldots, \lambda_q$

Формулировка подхода через минимальную ошибку

- Введем полный ортонормированный набор d-мерных базисных векторов $\{\mathbf{u}_i\}_{i=1}^d$, такой, что

$$\mathbf{u}_i^\top \mathbf{u}_j = \delta_{ij}$$

- Таким образом, для любого $\mathbf{x}_i$ выполняется: $\mathbf{x}_i = \sum_{j=1}^d \alpha_{ij}\mathbf{u}_j$
- Из-за ортонормированности получаем, что $\alpha_{ij} = \mathbf{x}_i^\top \mathbf{u}_j$, т.е.

$$\mathbf{x}_i = \sum_{j=1}^d (\mathbf{x}_i^\top \mathbf{u}_j)\mathbf{u}_j$$

- q-мерное линейное подпространство представляется первыми q базисных векторов, поэтому аппроксимация $\mathbf{x}_i$ равна

$$\widetilde{\mathbf{x}}_i = \sum_{j=1}^q z_{ij}\mathbf{u}_j + \sum_{j=q+1}^d b_j\mathbf{u}_j,$$

где $\{b_j\}$ одинаковые для всех точек данных константы

Формулировка подхода через минимальную ошибку

- Введем полный ортонормированный набор d-мерных базисных векторов $\{\mathbf{u}_i\}_{i=1}^d$, такой, что

$$\mathbf{u}_i^\top \mathbf{u}_j = \delta_{ij}$$

- Таким образом, для любого $\mathbf{x}_i$ выполняется: $\mathbf{x}_i = \sum_{j=1}^d \alpha_{ij}\mathbf{u}_j$
- Из-за ортонормированности получаем, что $\alpha_{ij} = \mathbf{x}_i^\top \mathbf{u}_j$, т.е.

$$\mathbf{x}_i = \sum_{j=1}^d (\mathbf{x}_i^\top \mathbf{u}_j)\mathbf{u}_j$$

- q-мерное линейное подпространство представляется первыми q базисных векторов, поэтому аппроксимация $\mathbf{x}_i$ равна

$$\widetilde{\mathbf{x}}_i = \sum_{j=1}^q z_{ij}\mathbf{u}_j + \sum_{j=q+1}^d b_j\mathbf{u}_j,$$

где $\{b_j\}$ одинаковые для всех точек данных константы

Формулировка подхода через минимальную ошибку

- Введем полный ортонормированный набор d-мерных базисных векторов $\{\mathbf{u}_i\}_{i=1}^d$, такой, что

$$\mathbf{u}_i^\top \mathbf{u}_j = \delta_{ij}$$

- Таким образом, для любого $\mathbf{x}_i$ выполняется: $\mathbf{x}_i = \sum_{j=1}^d \alpha_{ij}\mathbf{u}_j$
- Из-за ортонормированности получаем, что $\alpha_{ij} = \mathbf{x}_i^\top \mathbf{u}_j$, т.е.

$$\mathbf{x}_i = \sum_{j=1}^d (\mathbf{x}_i^\top \mathbf{u}_j)\mathbf{u}_j$$

- q-мерное линейное подпространство представляется первыми q базисных векторов, поэтому аппроксимация $\mathbf{x}_i$ равна

$$\widetilde{\mathbf{x}}_i = \sum_{j=1}^q z_{ij}\mathbf{u}_j + \sum_{j=q+1}^d b_j\mathbf{u}_j,$$

где $\{b_j\}$ одинаковые для всех точек данных константы

Формулировка подхода через минимальную ошибку

- Введем полный ортонормированный набор d-мерных базисных векторов $\{\mathbf{u}_i\}_{i=1}^d$, такой, что

$$\mathbf{u}_i^\top \mathbf{u}_j = \delta_{ij}$$

- Таким образом, для любого $\mathbf{x}_i$ выполняется: $\mathbf{x}_i = \sum_{j=1}^d \alpha_{ij}\mathbf{u}_j$
- Из-за ортонормированности получаем, что $\alpha_{ij} = \mathbf{x}_i^\top \mathbf{u}_j$, т.е.

$$\mathbf{x}_i = \sum_{j=1}^d (\mathbf{x}_i^\top \mathbf{u}_j)\mathbf{u}_j$$

- q-мерное линейное подпространство представляется первыми q базисных векторов, поэтому аппроксимация $\mathbf{x}_i$ равна

$$\widetilde{\mathbf{x}}_i = \sum_{j=1}^q z_{ij}\mathbf{u}_j + \sum_{j=q+1}^d b_j\mathbf{u}_j,$$

где $\{b_j\}$ одинаковые для всех точек данных константы

Формулировка подхода через минимальную ошибку

- Мера искажения

$$J = \frac{1}{m}\sum_{i=1}^{m}\|\mathbf{x}_i - \widetilde{\mathbf{x}}_i\|^2$$

- Приравнивая производные к нулю, мы получим

$$\{z_{ij} = \mathbf{x}_i^\top \mathbf{u}_j\}_{j=1}^{q},\ \{b_j = \overline{\mathbf{x}}^\top \mathbf{u}_j\}_{j=q+1}^{d}$$

- Поскольку $\mathbf{x}_i - \widetilde{\mathbf{x}}_i = \sum_{j=q+1}^{d}\{(\mathbf{x}_i - \overline{\mathbf{x}})^\top \mathbf{u}_j\}\mathbf{u}_j$, тогда

$$J = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=q+1}^{d}(\mathbf{x}_i^\top \mathbf{u}_j - \overline{\mathbf{x}}^\top \mathbf{u}_j)^2 = \sum_{j=q+1}^{d}\mathbf{u}_j^\top \mathbf{S}\mathbf{u}_j$$

Формулировка подхода через минимальную ошибку

- Мера искажения

$$J = \frac{1}{m}\sum_{i=1}^{m}\|\mathbf{x}_i - \widetilde{\mathbf{x}}_i\|^2$$

- Приравнивая производные к нулю, мы получим

$$\{z_{ij} = \mathbf{x}_i^\top \mathbf{u}_j\}_{j=1}^{q},\ \{b_j = \overline{\mathbf{x}}^\top \mathbf{u}_j\}_{j=q+1}^{d}$$

- Поскольку $\mathbf{x}_i - \widetilde{\mathbf{x}}_i = \sum_{j=q+1}^{d}\{(\mathbf{x}_i - \overline{\mathbf{x}})^\top \mathbf{u}_j\}\mathbf{u}_j$, тогда

$$J = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=q+1}^{d}(\mathbf{x}_i^\top \mathbf{u}_j - \overline{\mathbf{x}}^\top \mathbf{u}_j)^2 = \sum_{j=q+1}^{d}\mathbf{u}_j^\top \mathbf{S}\mathbf{u}_j$$

- Мера искажения

$$J = \frac{1}{m}\sum_{i=1}^{m}\|\mathbf{x}_i - \widetilde{\mathbf{x}}_i\|^2$$

- Приравнивая производные к нулю, мы получим

$$\{z_{ij} = \mathbf{x}_i^\top \mathbf{u}_j\}_{j=1}^{q},\ \{b_j = \overline{\mathbf{x}}^\top \mathbf{u}_j\}_{j=q+1}^{d}$$

- Поскольку $\mathbf{x}_i - \widetilde{\mathbf{x}}_i = \sum_{j=q+1}^{d}\{(\mathbf{x}_i - \overline{\mathbf{x}})^\top \mathbf{u}_j\}\mathbf{u}_j$, тогда

$$J = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=q+1}^{d}(\mathbf{x}_i^\top \mathbf{u}_j - \overline{\mathbf{x}}^\top \mathbf{u}_j)^2 = \sum_{j=q+1}^{d}\mathbf{u}_j^\top \mathbf{S}\mathbf{u}_j$$

- Например, в случае $d = 2$: путем минимизации

$$J = \mathbf{u}_2^\top \mathbf{S}\mathbf{u}_2 + \lambda_2(1 - \mathbf{u}_2^\top \mathbf{u}_2)$$

мы получаем это

$$\mathbf{S}\mathbf{u}_2 = \lambda_2 \mathbf{u}_2,\ J = \lambda_2,$$

то есть мы должны выбрать главное подпространство, которое будет связано с собственным вектором, имеющим большее собственное значение

- В общем случае $\{\mathbf{u}_i\}_{i=1}^q$ являются собственными векторами $\mathbf{S}\mathbf{u}_i = \lambda_i \mathbf{u}_i$ и

$$J = \sum_{i=q+1}^{d} \lambda_i$$

- Например, в случае $d = 2$: путем минимизации

$$J = \mathbf{u}_2^\top \mathbf{S}\mathbf{u}_2 + \lambda_2(1 - \mathbf{u}_2^\top \mathbf{u}_2)$$

мы получаем это

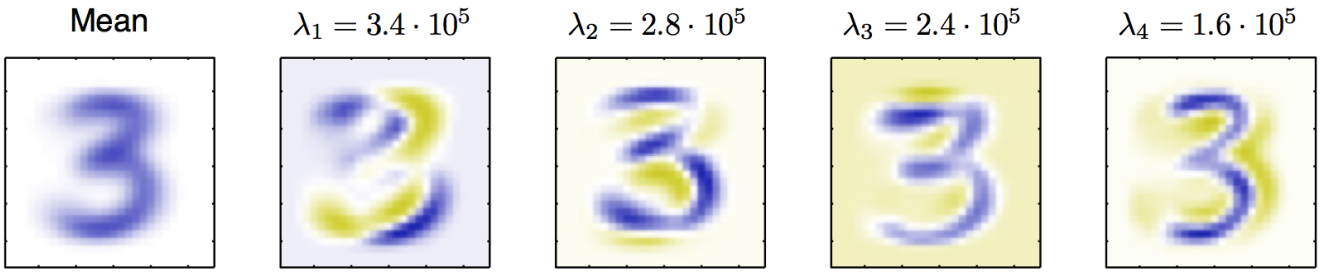
$$\mathbf{S}\mathbf{u}_2 = \lambda_2 \mathbf{u}_2,\ J = \lambda_2,$$

то есть мы должны выбрать главное подпространство, которое будет связано с собственным вектором, имеющим большее собственное значение

- В общем случае $\{\mathbf{u}_i\}_{i=1}^q$ являются собственными векторами $\mathbf{S}\mathbf{u}_i = \lambda_i \mathbf{u}_i$ и

$$J = \sum_{i=q+1}^{d} \lambda_i$$

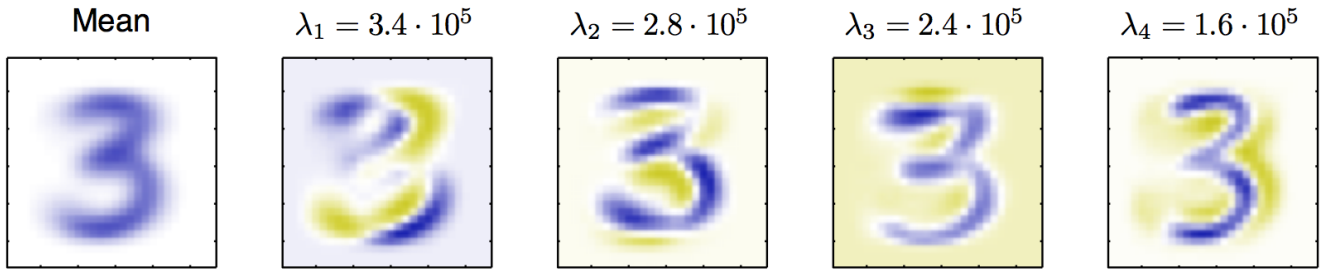
Приложения метода главных компонент (PCA)



- PCA приближение к вектору данных $\mathbf{x}_n$

$$\widetilde{\mathbf{x}}_i = \sum_{j=1}^{q} (\mathbf{x}_i^\top \mathbf{u}_j)\mathbf{u}_j + \sum_{j=q+1}^{d} (\overline{\mathbf{x}}^\top \mathbf{u}_j)\mathbf{u}_j$$

Приложения метода главных компонент (PCA)

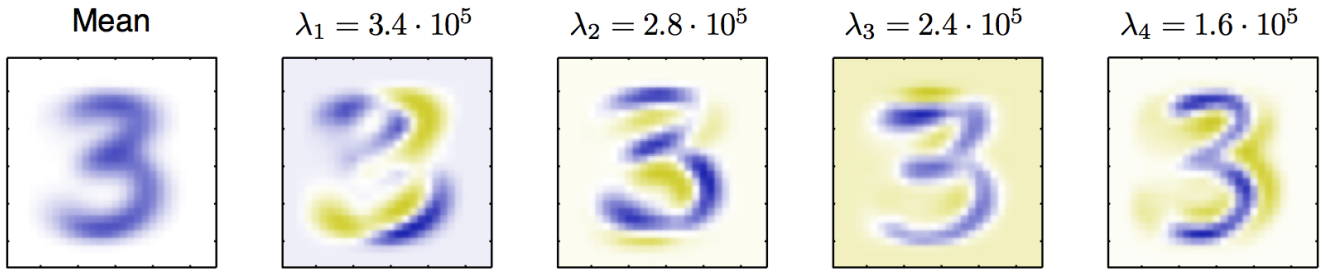


- PCA приближение к вектору данных $\mathbf{x}_n$

$$\widetilde{\mathbf{x}}_i = \sum_{j=1}^{q} (\mathbf{x}_i^\top \mathbf{u}_j)\mathbf{u}_j + \sum_{j=q+1}^{d} (\overline{\mathbf{x}}^\top \mathbf{u}_j)\mathbf{u}_j$$

$$= \overline{\mathbf{x}} + \sum_{j=1}^{q} (\mathbf{x}_j^\top - \overline{\mathbf{x}}^\top \mathbf{u}_j)\mathbf{u}_j,$$

Приложения метода главных компонент (PCA)



- PCA приближение к вектору данных $\mathbf{x}_n$

$$\widetilde{\mathbf{x}}_i = \sum_{j=1}^{q} (\mathbf{x}_i^\top \mathbf{u}_j)\mathbf{u}_j + \sum_{j=q+1}^{d} (\overline{\mathbf{x}}^\top \mathbf{u}_j)\mathbf{u}_j$$

$$= \overline{\mathbf{x}} + \sum_{j=1}^{q} (\mathbf{x}_j^\top - \overline{\mathbf{x}}^\top \mathbf{u}_j)\mathbf{u}_j,$$

где мы использовали равенство $\overline{\mathbf{x}} = \sum_{i=1}^{d} (\overline{\mathbf{x}}^\top \mathbf{u}_i)\mathbf{u}_i$

Приложения метода главных компонент (PCA)

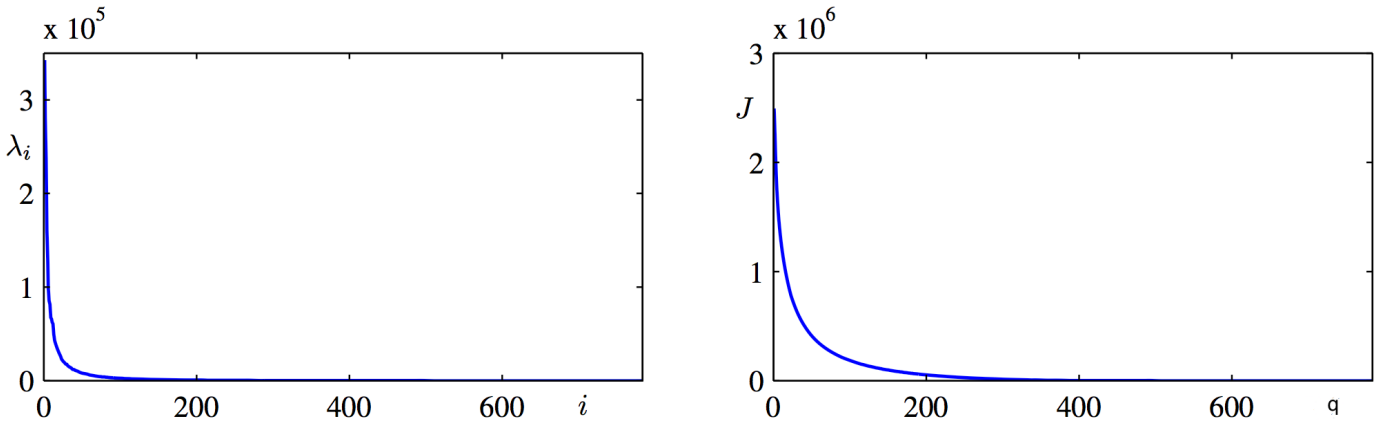


Рис. – Спектр собственных значений (слева). Сумма отброшенных собственных значений (справа)

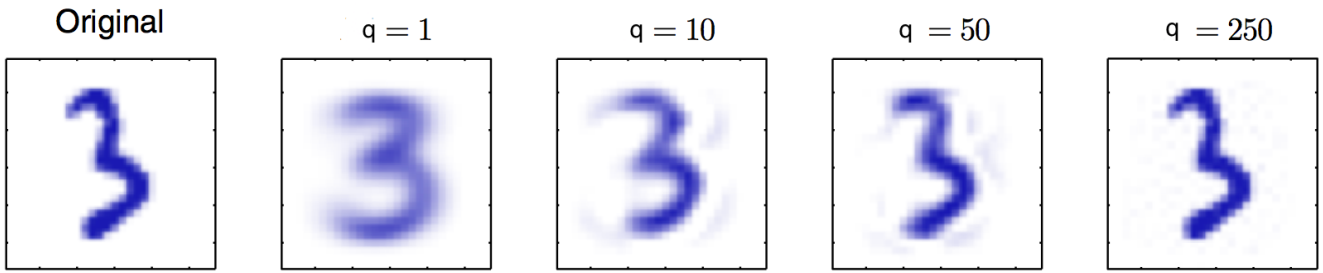


Рис. – PCA восстановление данных, состоящих из набора цифр. $q = d = 28 \times 28 = 784$ — это уже идеальная реконструкция

Преимущества вероятностного PCA

- Вероятностный PCA представляет собой ограниченную форму гауссовского распределения
- Обеспечивает EM-алгоритм для PCA: вычислительно эффективен, так как мы можем рассчитывать только необходимые компоненты
- Вероятностная модель + EM служит для устранения пропущенных значений
- Смеси вероятностных моделей PCA могут быть сформулированы с вероятностной точки зрения, что позволит использовать EM алгоритм
- Существование вероятностной функции $\Rightarrow$ прямое сравнение с другими вероятностными моделями плотности
- Вероятностный PCA может быть использован для моделирования класса условных плотностей
- Вероятностная модель PCA может быть использована для генерации образцов из распределения

Преимущества вероятностного PCA

- Вероятностный PCA представляет собой ограниченную форму гауссовского распределения
- Обеспечивает EM-алгоритм для PCA: вычислительно эффективен, так как мы можем рассчитывать только необходимые компоненты
- Вероятностная модель + EM служит для устранения пропущенных значений
- Смеси вероятностных моделей PCA могут быть сформулированы с вероятностной точки зрения, что позволит использовать EM алгоритм
- Существование вероятностной функции $\Rightarrow$ прямое сравнение с другими вероятностными моделями плотности
- Вероятностный PCA может быть использован для моделирования класса условных плотностей
- Вероятностная модель PCA может быть использована для генерации образцов из распределения

Преимущества вероятностного PCA

- Вероятностный PCA представляет собой ограниченную форму гауссовского распределения
- Обеспечивает EM-алгоритм для PCA: вычислительно эффективен, так как мы можем рассчитывать только необходимые компоненты
- Вероятностная модель + EM служит для устранения пропущенных значений
- Смеси вероятностных моделей PCA могут быть сформулированы с вероятностной точки зрения, что позволит использовать EM алгоритм
- Существование вероятностной функции $\Rightarrow$ прямое сравнение с другими вероятностными моделями плотности
- Вероятностный PCA может быть использован для моделирования класса условных плотностей
- Вероятностная модель PCA может быть использована для генерации образцов из распределения

Преимущества вероятностного PCA

- Вероятностный PCA представляет собой ограниченную форму гауссовского распределения
- Обеспечивает EM-алгоритм для PCA: вычислительно эффективен, так как мы можем рассчитывать только необходимые компоненты
- Вероятностная модель + EM служит для устранения пропущенных значений
- Смеси вероятностных моделей PCA могут быть сформулированы с вероятностной точки зрения, что позволит использовать EM алгоритм
- Существование вероятностной функции $\Rightarrow$ прямое сравнение с другими вероятностными моделями плотности
- Вероятностный PCA может быть использован для моделирования класса условных плотностей
- Вероятностная модель PCA может быть использована для генерации образцов из распределения

Преимущества вероятностного PCA

- Вероятностный PCA представляет собой ограниченную форму гауссовского распределения
- Обеспечивает EM-алгоритм для PCA: вычислительно эффективен, так как мы можем рассчитывать только необходимые компоненты
- Вероятностная модель + EM служит для устранения пропущенных значений
- Смеси вероятностных моделей PCA могут быть сформулированы с вероятностной точки зрения, что позволит использовать EM алгоритм
- Существование вероятностной функции $\Rightarrow$ прямое сравнение с другими вероятностными моделями плотности
- Вероятностный PCA может быть использован для моделирования класса условных плотностей
- Вероятностная модель PCA может быть использована для генерации образцов из распределения

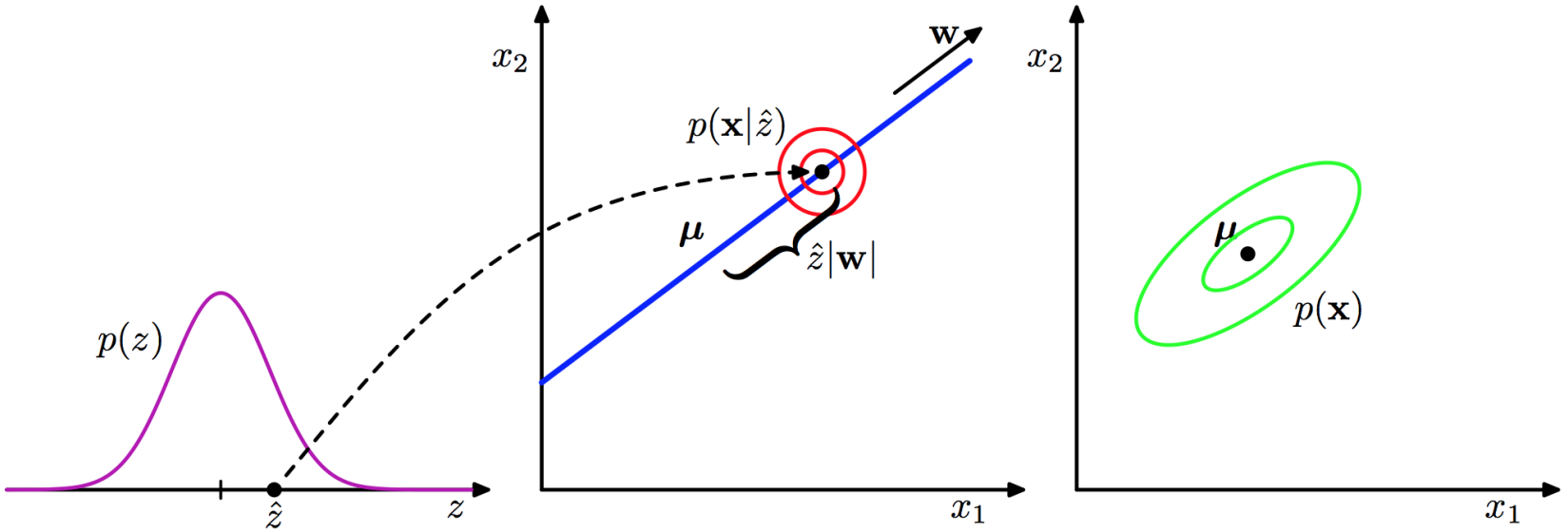
Преимущества вероятностного PCA

- Вероятностный PCA представляет собой ограниченную форму гауссовского распределения
- Обеспечивает EM-алгоритм для PCA: вычислительно эффективен, так как мы можем рассчитывать только необходимые компоненты
- Вероятностная модель + EM служит для устранения пропущенных значений
- Смеси вероятностных моделей PCA могут быть сформулированы с вероятностной точки зрения, что позволит использовать EM алгоритм
- Существование вероятностной функции $\Rightarrow$ прямое сравнение с другими вероятностными моделями плотности
- Вероятностный PCA может быть использован для моделирования класса условных плотностей
- Вероятностная модель PCA может быть использована для генерации образцов из распределения

Преимущества вероятностного PCA

- Вероятностный PCA представляет собой ограниченную форму гауссовского распределения
- Обеспечивает EM-алгоритм для PCA: вычислительно эффективен, так как мы можем рассчитывать только необходимые компоненты
- Вероятностная модель + EM служит для устранения пропущенных значений
- Смеси вероятностных моделей PCA могут быть сформулированы с вероятностной точки зрения, что позволит использовать EM алгоритм
- Существование вероятностной функции $\Rightarrow$ прямое сравнение с другими вероятностными моделями плотности
- Вероятностный PCA может быть использован для моделирования класса условных плотностей
- Вероятностная модель PCA может быть использована для генерации образцов из распределения

Вероятностный PCA

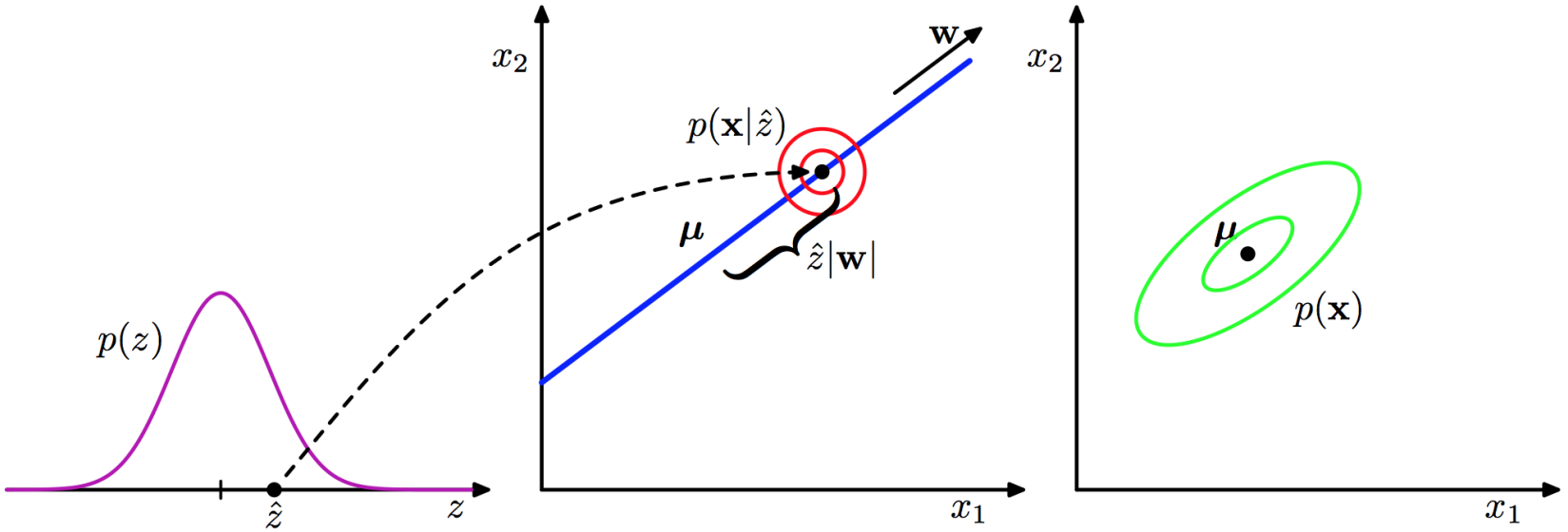


- Предположим, что $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$, $\mathbf{z} \in \mathbb{R}^q$ $(q < d)$
- По аналогии

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I}), \text{ i.e. } \mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \; \mathbf{x} \in \mathbb{R}^d,$$

где $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \sigma^2\mathbf{I})$

Вероятностный PCA



- Предположим, что $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$, $\mathbf{z} \in \mathbb{R}^q$ $(q < d)$
- По аналогии

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I}), \text{ i.e. } \mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}, \ \mathbf{x} \in \mathbb{R}^d,$$

где $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \sigma^2\mathbf{I})$

- Нам необходимо определить $\mathbf{W}$ и σ^2. Таким образом, нам нужно маргинальное распределение $p(\mathbf{x})$

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

- Получаем, что $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C})$, где

$$\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$$

- В этой параметризации есть избыточность, соответствующая вращениям латентных пространственных координат $\widetilde{\mathbf{W}} = \mathbf{W}\mathbf{R}$, где $\mathbf{R}$ повороты скрытых координат пространства: для

$$\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^\top = \mathbf{W}\mathbf{R}\mathbf{R}^\top\mathbf{W}^\top = \mathbf{W}\mathbf{W}^\top$$

- Нам необходимо определить $\mathbf{W}$ и σ^2. Таким образом, нам нужно маргинальное распределение $p(\mathbf{x})$

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

- Получаем, что $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C})$, где

$$\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$$

- В этой параметризации есть избыточность, соответствующая вращениям латентных пространственных координат $\widetilde{\mathbf{W}} = \mathbf{W}\mathbf{R}$, где $\mathbf{R}$ повороты скрытых координат пространства: для

$$\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^\top = \mathbf{W}\mathbf{R}\mathbf{R}^\top\mathbf{W}^\top = \mathbf{W}\mathbf{W}^\top$$

- Нам необходимо определить $\mathbf{W}$ и σ^2. Таким образом, нам нужно маргинальное распределение $p(\mathbf{x})$

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$$

- Получаем, что $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{C})$, где

$$\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$$

- В этой параметризации есть избыточность, соответствующая вращениям латентных пространственных координат $\widetilde{\mathbf{W}} = \mathbf{W}\mathbf{R}$, где $\mathbf{R}$ повороты скрытых координат пространства: для

$$\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^\top = \mathbf{W}\mathbf{R}\mathbf{R}^\top\mathbf{W}^\top = \mathbf{W}\mathbf{W}^\top$$

- Обращение матрицы $\mathbf{C}$ размера $d \times d$:

$$\mathbf{C}^{-1} = \sigma^{-1}\mathbf{I} - \sigma^{-2}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{\top},$$

где матрица $\mathbf{M}$ размера $q \times q$ имеет вид

$$\mathbf{M} = \mathbf{W}^{\top}\mathbf{W} + \sigma^{2}\mathbf{I}$$

- Таким образом, стоимость обращения $\mathbf{C}$ уменьшается с $O(d^3)$ до $O(q^3)$
- Апостериорное распределение $p(\mathbf{z}|\mathbf{x})$

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mathbf{M}^{-1}\mathbf{W}^{\top}(\mathbf{x} - \boldsymbol{\mu}), \sigma^{-2}\mathbf{M})$$

- Обращение матрицы $\mathbf{C}$ размера $d \times d$:

$$\mathbf{C}^{-1} = \sigma^{-1}\mathbf{I} - \sigma^{-2}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{\top},$$

  где матрица $\mathbf{M}$ размера $q \times q$ имеет вид

$$\mathbf{M} = \mathbf{W}^{\top}\mathbf{W} + \sigma^{2}\mathbf{I}$$

- Таким образом, стоимость обращения $\mathbf{C}$ уменьшается с $O(d^3)$ до $O(q^3)$
- Апостериорное распределение $p(\mathbf{z}|\mathbf{x})$

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mathbf{M}^{-1}\mathbf{W}^{\top}(\mathbf{x} - \boldsymbol{\mu}), \sigma^{-2}\mathbf{M})$$

- Обращение матрицы $\mathbf{C}$ размера $d \times d$:

$$\mathbf{C}^{-1} = \sigma^{-1}\mathbf{I} - \sigma^{-2}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{\top},$$

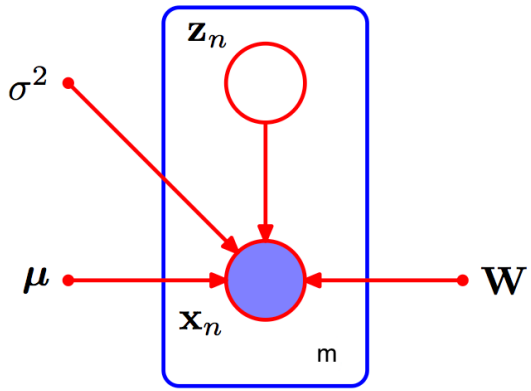
  где матрица $\mathbf{M}$ размера $q \times q$ имеет вид

$$\mathbf{M} = \mathbf{W}^{\top}\mathbf{W} + \sigma^{2}\mathbf{I}$$

- Таким образом, стоимость обращения $\mathbf{C}$ уменьшается с $O(d^3)$ до $O(q^3)$
- Апостериорное распределение $p(\mathbf{z}|\mathbf{x})$

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mathbf{M}^{-1}\mathbf{W}^{\top}(\mathbf{x} - \boldsymbol{\mu}), \sigma^{-2}\mathbf{M})$$

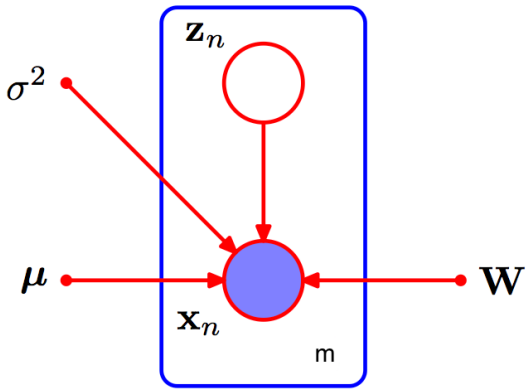
Метод максимального правдоподобия для PCA



- Учитывая набор данных $\mathbf{X}_m = \{\mathbf{x}_i\}_{i=1}^m$ логарифм правдоподобия

$$\log p(\mathbf{X}_m|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) = \sum_{i=1}^{m} \log p(\mathbf{x}_i|\mathbf{W}, \boldsymbol{\mu}, \sigma^2)$$

Метод максимального правдоподобия для PCA



- Учитывая набор данных $\mathbf{X}_m = \{\mathbf{x}_i\}_{i=1}^m$ логарифм правдоподобия

$$\log p(\mathbf{X}_m|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) = \sum_{i=1}^{m} \log p(\mathbf{x}_i|\mathbf{W}, \boldsymbol{\mu}, \sigma^2)$$

$$= -\frac{md}{2}\log(2\pi) - \frac{m}{2}\log|\mathbf{C}| - \frac{1}{2}\sum_{i=1}^{m}(\mathbf{x}_i - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$$

- Оптимизируя по $\boldsymbol{\mu}$, получим $\boldsymbol{\mu} = \overline{\mathbf{x}}$ и

$$\log p(\mathbf{X}_m|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) = -\frac{m}{2}\{d\log(2\pi) + \log|\mathbf{C}| + \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{S})\},$$

где **S** — ковариационная матрица данных

- ML для $\mathbf{W}$ и σ^2: все стационарные точки логарифмического правдоподобия имеют вид

$$\mathbf{W}_{ML} = \mathbf{U}_q(\mathbf{L}_q - \sigma^2_{ML}\mathbf{I})^{1/2}\mathbf{R},\ \sigma^2_{ML} = \frac{1}{d-q}\sum_{i=q+1}^{d}\lambda_i$$

где

— $\mathbf{U}_q \in \mathbb{R}^{d\times q}$ матрица, столбцы которой задаются любым подмножеством (размером q) собственных векторов ковариационной матрицы данных **S**,

— $\mathbf{L}_q$ диагональная матрица $q \times q$ с элементами λ_i,

— $\mathbf{R}$ произвольная ортогональная матрица размером $q \times q$.

Метод максимального правдоподобия для PCA

- Оптимизируя по $\boldsymbol{\mu}$, получим $\boldsymbol{\mu} = \overline{\mathbf{x}}$ и

$$\log p(\mathbf{X}_m|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) = -\frac{m}{2}\{d\log(2\pi) + \log|\mathbf{C}| + \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{S})\},$$

где $\mathbf{S}$ — ковариационная матрица данных

- ML для $\mathbf{W}$ и σ^2: все стационарные точки логарифмического правдоподобия имеют вид

$$\mathbf{W}_{ML} = \mathbf{U}_q(\mathbf{L}_q - \sigma^2_{ML}\mathbf{I})^{1/2}\mathbf{R},\ \sigma^2_{ML} = \frac{1}{d-q}\sum_{i=q+1}^{d}\lambda_i$$

где

  - $\mathbf{U}_q \in \mathbb{R}^{d\times q}$ матрица, столбцы которой задаются любым подмножеством (размером q) собственных векторов ковариационной матрицы данных $\mathbf{S}$,
  - $\mathbf{L}_q$ диагональная матрица $q \times q$ с элементами λ_i,
  - $\mathbf{R}$ произвольная ортогональная матрица размером $q \times q$.

- Оптимизируя по $\boldsymbol{\mu}$, получим $\boldsymbol{\mu} = \overline{\mathbf{x}}$ и

$$\log p(\mathbf{X}_m|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) = -\frac{m}{2}\{d\log(2\pi) + \log|\mathbf{C}| + \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{S})\},$$

  где $\mathbf{S}$ — ковариационная матрица данных

- ML для $\mathbf{W}$ и σ^2: все стационарные точки логарифмического правдоподобия имеют вид

$$\mathbf{W}_{ML} = \mathbf{U}_q(\mathbf{L}_q - \sigma^2_{ML}\mathbf{I})^{1/2}\mathbf{R},\ \sigma^2_{ML} = \frac{1}{d-q}\sum_{i=q+1}^{d}\lambda_i$$

  где
  - $\mathbf{U}_q \in \mathbb{R}^{d\times q}$ матрица, столбцы которой задаются любым подмножеством (размером q) собственных векторов ковариационной матрицы данных $\mathbf{S}$,
  - $\mathbf{L}_q$ диагональная матрица $q \times q$ с элементами λ_i,
  - $\mathbf{R}$ произвольная ортогональная матрица размером $q \times q$.

- Оптимизируя по $\boldsymbol{\mu}$, получим $\boldsymbol{\mu} = \overline{\mathbf{x}}$ и

$$\log p(\mathbf{X}_m|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) = -\frac{m}{2}\{d\log(2\pi) + \log|\mathbf{C}| + \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{S})\},$$

  где $\mathbf{S}$ — ковариационная матрица данных

- ML для $\mathbf{W}$ и σ^2: все стационарные точки логарифмического правдоподобия имеют вид

$$\mathbf{W}_{ML} = \mathbf{U}_q(\mathbf{L}_q - \sigma^2_{ML}\mathbf{I})^{1/2}\mathbf{R},\ \sigma^2_{ML} = \frac{1}{d-q}\sum_{i=q+1}^{d}\lambda_i$$

  где
  - $\mathbf{U}_q \in \mathbb{R}^{d\times q}$ матрица, столбцы которой задаются любым подмножеством (размером q) собственных векторов ковариационной матрицы данных $\mathbf{S}$,
  - $\mathbf{L}_q$ диагональная матрица $q \times q$ с элементами λ_i,
  - $\mathbf{R}$ произвольная ортогональная матрица размером $q \times q$.

- Оптимизируя по $\boldsymbol{\mu}$, получим $\boldsymbol{\mu} = \overline{\mathbf{x}}$ и

$$\log p(\mathbf{X}_m|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) = -\frac{m}{2}\{d\log(2\pi) + \log|\mathbf{C}| + \mathrm{Tr}(\mathbf{C}^{-1}\mathbf{S})\},$$

  где $\mathbf{S}$ — ковариационная матрица данных
- ML для $\mathbf{W}$ и σ^2: все стационарные точки логарифмического правдоподобия имеют вид

$$\mathbf{W}_{ML} = \mathbf{U}_q(\mathbf{L}_q - \sigma^2_{ML}\mathbf{I})^{1/2}\mathbf{R},\ \sigma^2_{ML} = \frac{1}{d-q}\sum_{i=q+1}^{d}\lambda_i$$

  где
  - $\mathbf{U}_q \in \mathbb{R}^{d\times q}$ матрица, столбцы которой задаются любым подмножеством (размером q) собственных векторов ковариационной матрицы данных $\mathbf{S}$,
  - $\mathbf{L}_q$ диагональная матрица $q \times q$ с элементами λ_i,
  - $\mathbf{R}$ произвольная ортогональная матрица размером $q \times q$.

Метод максимального правдоподобия для PCA

- Для безусловного распределения $p(\mathbf{x})$ получаем, что

$$\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}] = \boldsymbol{\mu}$$
$$\text{cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon})(\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon})^\top] = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I} = \mathbf{C}$$

- Таким образом, $\mathbf{C}$ не зависит от $\mathbf{R}$ для

$$\mathbf{W}_{ML} = \mathbf{U}_q(\mathbf{L}_q - \sigma^2_{ML}\mathbf{I})^{1/2}\mathbf{R}$$

- Если $\mathbf{v}$ ортогонально главному подпространству, то $\mathbf{v}^\top\mathbf{U} = \mathbf{0}$, т.е. $\mathbf{v}^\top\mathbf{C}\mathbf{v} = \sigma^2$
- Если $\mathbf{v} = \mathbf{u}_i$, то $\mathbf{v}^\top\mathbf{C}\mathbf{v} = (\lambda_i - \sigma^2) + \sigma^2 = \lambda_i$
- Для $\mathbf{R} = \mathbf{I}$ мы получаем обычный PCA, иначе столбцы $\mathbf{W}$ должны быть не ортогональными

Метод максимального правдоподобия для PCA

- Для безусловного распределения $p(\mathbf{x})$ получаем, что

$$\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}] = \boldsymbol{\mu}$$
$$\text{cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon})(\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon})^\top] = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I} = \mathbf{C}$$

- Таким образом, $\mathbf{C}$ не зависит от $\mathbf{R}$ для

$$\mathbf{W}_{ML} = \mathbf{U}_q(\mathbf{L}_q - \sigma^2_{ML}\mathbf{I})^{1/2}\mathbf{R}$$

- Если $\mathbf{v}$ ортогонально главному подпространству, то $\mathbf{v}^\top\mathbf{U} = \mathbf{0}$, т.е. $\mathbf{v}^\top\mathbf{C}\mathbf{v} = \sigma^2$
- Если $\mathbf{v} = \mathbf{u}_i$, то $\mathbf{v}^\top\mathbf{C}\mathbf{v} = (\lambda_i - \sigma^2) + \sigma^2 = \lambda_i$
- Для $\mathbf{R} = \mathbf{I}$ мы получаем обычный PCA, иначе столбцы $\mathbf{W}$ должны быть не ортогональными

Метод максимального правдоподобия для PCA

- Для безусловного распределения $p(\mathbf{x})$ получаем, что

$$\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}] = \boldsymbol{\mu}$$
$$\text{cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon})(\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon})^\top] = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I} = \mathbf{C}$$

- Таким образом, $\mathbf{C}$ не зависит от $\mathbf{R}$ для

$$\mathbf{W}_{ML} = \mathbf{U}_q(\mathbf{L}_q - \sigma^2_{ML}\mathbf{I})^{1/2}\mathbf{R}$$

- Если $\mathbf{v}$ ортогонально главному подпространству, то $\mathbf{v}^\top\mathbf{U} = \mathbf{0}$, т.е. $\mathbf{v}^\top\mathbf{C}\mathbf{v} = \sigma^2$
- Если $\mathbf{v} = \mathbf{u}_i$, то $\mathbf{v}^\top\mathbf{C}\mathbf{v} = (\lambda_i - \sigma^2) + \sigma^2 = \lambda_i$
- Для $\mathbf{R} = \mathbf{I}$ мы получаем обычный PCA, иначе столбцы $\mathbf{W}$ должны быть не ортогональными

Метод максимального правдоподобия для PCA

- Для безусловного распределения $p(\mathbf{x})$ получаем, что

$$\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}] = \boldsymbol{\mu}$$
$$\text{cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon})(\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon})^\top] = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I} = \mathbf{C}$$

- Таким образом, $\mathbf{C}$ не зависит от $\mathbf{R}$ для

$$\mathbf{W}_{ML} = \mathbf{U}_q(\mathbf{L}_q - \sigma^2_{ML}\mathbf{I})^{1/2}\mathbf{R}$$

- Если $\mathbf{v}$ ортогонально главному подпространству, то $\mathbf{v}^\top\mathbf{U} = \mathbf{0}$, т.е. $\mathbf{v}^\top\mathbf{C}\mathbf{v} = \sigma^2$
- Если $\mathbf{v} = \mathbf{u}_i$, то $\mathbf{v}^\top\mathbf{C}\mathbf{v} = (\lambda_i - \sigma^2) + \sigma^2 = \lambda_i$
- Для $\mathbf{R} = \mathbf{I}$ мы получаем обычный PCA, иначе столбцы $\mathbf{W}$ должны быть не ортогональными

Метод максимального правдоподобия для PCA

- Для безусловного распределения $p(\mathbf{x})$ получаем, что

$$\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}] = \boldsymbol{\mu}$$
$$\text{cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon})(\mathbf{W}\mathbf{z} + \boldsymbol{\epsilon})^\top] = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I} = \mathbf{C}$$

- Таким образом, $\mathbf{C}$ не зависит от $\mathbf{R}$ для

$$\mathbf{W}_{ML} = \mathbf{U}_q(\mathbf{L}_q - \sigma^2_{ML}\mathbf{I})^{1/2}\mathbf{R}$$

- Если $\mathbf{v}$ ортогонально главному подпространству, то $\mathbf{v}^\top\mathbf{U} = \mathbf{0}$, т.е. $\mathbf{v}^\top\mathbf{C}\mathbf{v} = \sigma^2$
- Если $\mathbf{v} = \mathbf{u}_i$, то $\mathbf{v}^\top\mathbf{C}\mathbf{v} = (\lambda_i - \sigma^2) + \sigma^2 = \lambda_i$
- Для $\mathbf{R} = \mathbf{I}$ мы получаем обычный PCA, иначе столбцы $\mathbf{W}$ должны быть не ортогональными

Обычный PCA против Байесовского PCA

- Обычный PCA: проекция точек из d-мерного пространства данных на q-мерное линейное подпространство ($d > q$)
- Вероятностный PCA: отображение из скрытого пространства в пространство данных. Мы можем обратить это отображение, используя теорему Байеса (визуализация и сжатие данных)
- Среднее значение определяется

$$\mathbb{E}[\mathbf{z}|\mathbf{x}] = \mathbf{M}^{-1}\mathbf{W}_{ML}^{\top}(\mathbf{x} - \overline{\mathbf{x}})$$

- Апостприорная ковариация равна $\mathrm{cov}[\mathbf{z}] = \sigma^2\mathbf{M}^{-1}$

Обычный PCA против Байесовского PCA

- Обычный PCA: проекция точек из d-мерного пространства данных на q-мерное линейное подпространство ($d > q$)
- Вероятностный PCA: отображение из скрытого пространства в пространство данных. Мы можем обратить это отображение, используя теорему Байеса (визуализация и сжатие данных)
- Среднее значение определяется

$$\mathbb{E}[\mathbf{z}|\mathbf{x}] = \mathbf{M}^{-1}\mathbf{W}_{ML}^{\top}(\mathbf{x} - \overline{\mathbf{x}})$$

- Апостприорная ковариация равна $\text{cov}[\mathbf{z}] = \sigma^2\mathbf{M}^{-1}$

Обычный PCA против Байесовского PCA

- Обычный PCA: проекция точек из d-мерного пространства данных на q-мерное линейное подпространство ($d > q$)
- Вероятностный PCA: отображение из скрытого пространства в пространство данных. Мы можем обратить это отображение, используя теорему Байеса (визуализация и сжатие данных)
- Среднее значение определяется

$$\mathbb{E}[\mathbf{z}|\mathbf{x}] = \mathbf{M}^{-1}\mathbf{W}_{ML}^{\top}(\mathbf{x} - \overline{\mathbf{x}})$$

- Апостприорная ковариация равна $\mathrm{cov}[\mathbf{z}] = \sigma^2\mathbf{M}^{-1}$

Обычный PCA против Байесовского PCA

- Обычный PCA: проекция точек из d-мерного пространства данных на q-мерное линейное подпространство ($d > q$)
- Вероятностный PCA: отображение из скрытого пространства в пространство данных. Мы можем обратить это отображение, используя теорему Байеса (визуализация и сжатие данных)
- Среднее значение определяется

$$\mathbb{E}[\mathbf{z}|\mathbf{x}] = \mathbf{M}^{-1}\mathbf{W}_{ML}^{\top}(\mathbf{x} - \overline{\mathbf{x}})$$

- Апостприорная ковариация равна $\mathrm{cov}[\mathbf{z}] = \sigma^2\mathbf{M}^{-1}$

Обычный PCA против Байесовского PCA

- Обычное распределение Гаусса: $d(d+1)/2$ параметра.
- Вероятностный PCA: определить d-мерное Гауссовское распределения, сохраняя q наиболее значимых корреляций. Число степеней свободы в ковариационной матрице $\mathbf{C}$ определяется по формуле

$$dq + 1 - q(q-1)/2,$$

поскольку

  - $dq + 1$ for $\mathbf{W}$ and σ^2
  - -минус $q(q-1)/2$ параметров для $\mathbf{R}$ (избыточность в параметризации, связанная с вращениями)

Обычный PCA против Байесовского PCA

- Обычное распределение Гаусса: $d(d+1)/2$ параметра.
- Вероятностный PCA: определить d-мерное Гауссовское распределения, сохраняя q наиболее значимых корреляций. Число степеней свободы в ковариационной матрице $\mathbf{C}$ определяется по формуле

$$dq + 1 - q(q-1)/2,$$

поскольку

- $dq + 1$ for $\mathbf{W}$ and σ^2
- -минус $q(q-1)/2$ параметров для $\mathbf{R}$ (избыточность в параметризации, связанная с вращениями)

Обычный PCA против Байесовского PCA

- Обычное распределение Гаусса: $d(d+1)/2$ параметра.
- Вероятностный PCA: определить d-мерное Гауссовское распределения, сохраняя q наиболее значимых корреляций. Число степеней свободы в ковариационной матрице $\mathbf{C}$ определяется по формуле

$$dq + 1 - q(q-1)/2,$$

поскольку

- $dq + 1$ for $\mathbf{W}$ and σ^2
- -минус $q(q-1)/2$ параметров для $\mathbf{R}$ (избыточность в параметризации, связанная с вращениями)

Обычный PCA против Байесовского PCA

- Обычное распределение Гаусса: $d(d+1)/2$ параметра.
- Вероятностный PCA: определить d-мерное Гауссовское распределения, сохраняя q наиболее значимых корреляций. Число степеней свободы в ковариационной матрице $\mathbf{C}$ определяется по формуле

$$dq + 1 - q(q-1)/2,$$

поскольку

  - $dq + 1$ for $\mathbf{W}$ and σ^2
  - -минус $q(q-1)/2$ параметров для $\mathbf{R}$ (избыточность в параметризации, связанная с вращениями)

EM алгоритм для PCA

- Мы уже получили точное решение в явном виде для MLE. Зачем нам нужен EM-алгоритм?
- В пространствах с высокой размерностью могут быть вычислительные преимущества в использовании итерационной EM-процедуры, а не в непосредственной работе с выборочной ковариационной матрицей
- Общая структура EM
  - мы записываем логарифм полного правдоподобия
  - считаем его мат.ожидание с учётом апостериорного распределения латентных переменных со старыми параметрами
  - максимизируем ожидание полного правдоподобия, затем обновляем значения параметров

EM алгоритм для PCA

- Мы уже получили точное решение в явном виде для MLE. Зачем нам нужен EM-алгоритм?
- В пространствах с высокой размерностью могут быть вычислительные преимущества в использовании итерационной EM-процедуры, а не в непосредственной работе с выборочной ковариационной матрицей
- Общая структура EM
  - мы записываем логарифм полного правдоподобия
  - считаем его мат.ожидание с учётом апостериорного распределения латентных переменных со старыми параметрами
  - максимизируем ожидание полного правдоподобия, затем обновляем значения параметров

EM алгоритм для PCA

- Мы уже получили точное решение в явном виде для MLE. Зачем нам нужен EM-алгоритм?
- В пространствах с высокой размерностью могут быть вычислительные преимущества в использовании итерационной EM-процедуры, а не в непосредственной работе с выборочной ковариационной матрицей
- Общая структура EM
  - мы записываем логарифм полного правдоподобия
  - считаем его мат.ожидание с учётом апостериорного распределения латентных переменных со старыми параметрами
  - максимизируем ожидание полного правдоподобия, затем обновляем значения параметров

EM алгоритм для PCA

- Мы уже получили точное решение в явном виде для MLE. Зачем нам нужен EM-алгоритм?
- В пространствах с высокой размерностью могут быть вычислительные преимущества в использовании итерационной EM-процедуры, а не в непосредственной работе с выборочной ковариационной матрицей
- Общая структура EM
  - мы записываем логарифм полного правдоподобия
  - считаем его мат.ожидание с учётом апостериорного распределения латентных переменных со старыми параметрами
  - максимизируем ожидание полного правдоподобия, затем обновляем значения параметров

EM алгоритм для PCA

- Мы уже получили точное решение в явном виде для MLE. Зачем нам нужен EM-алгоритм?
- В пространствах с высокой размерностью могут быть вычислительные преимущества в использовании итерационной EM-процедуры, а не в непосредственной работе с выборочной ковариационной матрицей
- Общая структура EM
  - мы записываем логарифм полного правдоподобия
  - считаем его мат.ожидание с учётом апостериорного распределения латентных переменных со старыми параметрами
  - максимизируем ожидание полного правдоподобия, затем обновляем значения параметров

EM алгоритм для PCA

- Мы уже получили точное решение в явном виде для MLE. Зачем нам нужен EM-алгоритм?
- В пространствах с высокой размерностью могут быть вычислительные преимущества в использовании итерационной EM-процедуры, а не в непосредственной работе с выборочной ковариационной матрицей
- Общая структура EM
  - мы записываем логарифм полного правдоподобия
  - считаем его мат.ожидание с учётом апостериорного распределения латентных переменных со старыми параметрами
  - максимизируем ожидание полного правдоподобия, затем обновляем значения параметров

- Логарифм полного правдоподобия данных имеет вид

$$\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2) = \sum_{n=1}^{m} \{\log p(\mathbf{x}_n|\mathbf{z}_n) + \log p(\mathbf{z}_n)\}$$

- MLE-оценка для $\boldsymbol{\mu}$ равна $\overline{\mathbf{x}}$, таким образом, подставляя среднее значение выборки и выбирая ожидание в соответствии с апостериорным распределением по скрытым переменным

$$\mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2)] = -\sum_{n=1}^{m} \Big\{ \frac{d}{2} \log(2\pi\sigma^2) + \frac{1}{2} \mathrm{Tr}(\mathbb{E}[\mathbf{z}_n \mathbf{z}_n^\top]) + \frac{1}{2\sigma^2} \|\mathbf{x}_n - \boldsymbol{\mu}\|^2 - \frac{1}{\sigma^2} \mathbb{E}[\mathbf{z}_n]^\top \mathbf{W}^\top (\mathbf{x}_n - \boldsymbol{\mu}) + \frac{1}{2\sigma^2} \mathrm{Tr}(\mathbb{E}[\mathbf{z}_n \mathbf{z}_n^\top] \mathbf{W}^\top \mathbf{W}) \Big\}$$

- Логарифм полного правдоподобия данных имеет вид

$$\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2) = \sum_{n=1}^{m} \{\log p(\mathbf{x}_n|\mathbf{z}_n) + \log p(\mathbf{z}_n)\}$$

- MLE-оценка для $\boldsymbol{\mu}$ равна $\overline{\mathbf{x}}$, таким образом, подставляя среднее значение выборки и выбирая ожидание в соответствии с апостериорным распределением по скрытым переменным

$$\mathbb{E}[\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2)] = -\sum_{n=1}^{m} \Big\{ \frac{d}{2} \log(2\pi\sigma^2) + \frac{1}{2} \mathrm{Tr}(\mathbb{E}[\mathbf{z}_n \mathbf{z}_n^\top]) + \frac{1}{2\sigma^2} \|\mathbf{x}_n - \boldsymbol{\mu}\|^2 - \frac{1}{\sigma^2} \mathbb{E}[\mathbf{z}_n]^\top \mathbf{W}^\top (\mathbf{x}_n - \boldsymbol{\mu}) + \frac{1}{2\sigma^2} \mathrm{Tr}(\mathbb{E}[\mathbf{z}_n \mathbf{z}_n^\top] \mathbf{W}^\top \mathbf{W}) \Big\}$$

На шаге E мы используем старые значения параметров для оценки

$$\mathbb{E}[\mathbf{z}_n] = \mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{x}_n - \overline{\mathbf{x}})$$
$$\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top] = \mathrm{cov}[\mathbf{z}_n] + \mathbb{E}[\mathbf{z}_n]\mathbb{E}[\mathbf{z}_n]^\top$$

На шаге E мы используем старые значения параметров для оценки

$$\mathbb{E}[\mathbf{z}_n] = \mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{x}_n - \overline{\mathbf{x}})$$
$$\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top] = \mathrm{cov}[\mathbf{z}_n] + \mathbb{E}[\mathbf{z}_n]\mathbb{E}[\mathbf{z}_n]^\top$$

На шаге M максимизацией получаем $\mathbf{W}$ и σ^2:

На шаге E мы используем старые значения параметров для оценки

$$\mathbb{E}[\mathbf{z}_n] = \mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{x}_n - \overline{\mathbf{x}})$$
$$\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top] = \operatorname{cov}[\mathbf{z}_n] + \mathbb{E}[\mathbf{z}_n]\mathbb{E}[\mathbf{z}_n]^\top$$

На шаге M максимизацией получаем $\mathbf{W}$ и σ^2:

$$\mathbf{W}_{\text{new}} = \left[\sum_{n=1}^{m}(\mathbf{x}_n - \overline{\mathbf{x}})\mathbb{E}[\mathbf{z}_n]^\top\right]\left[\sum_{n=1}^{m}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top]\right]^{-1}$$

На шаге E мы используем старые значения параметров для оценки

$$\mathbb{E}[\mathbf{z}_n] = \mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{x}_n - \overline{\mathbf{x}})$$
$$\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top] = \operatorname{cov}[\mathbf{z}_n] + \mathbb{E}[\mathbf{z}_n]\mathbb{E}[\mathbf{z}_n]^\top$$

На шаге M максимизацией получаем $\mathbf{W}$ и σ^2:

$$\mathbf{W}_{\text{new}} = \left[\sum_{n=1}^{m}(\mathbf{x}_n - \overline{\mathbf{x}})\mathbb{E}[\mathbf{z}_n]^\top\right]\left[\sum_{n=1}^{m}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top]\right]^{-1}$$

$$\sigma^2_{\text{new}} = \frac{1}{md}\sum_{n=1}^{m}\Big\{\|\mathbf{x}_n - \overline{\mathbf{x}}\|^2 - 2\mathbb{E}[\mathbf{z}_n]^\top\mathbf{W}_{\text{new}}^\top(\mathbf{x}_n - \overline{\mathbf{x}}) + \operatorname{Tr}\left(\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top]\mathbf{W}_{\text{new}}^\top\mathbf{W}_{\text{new}}\right)$$

EM против MLE

- Преимущество итеративного алгоритма EM для PCA: вычислительная эффективность для крупномасштабных приложений
- PCA: $O(d^3)$ для собственного разложения или $O(qd^2)$, если нам нужны первые q собственных векторов
- Однако, нам нужно $O(md^2)$ для вычисления ковариационной матрицы.
- В случае алгоритма EM нам нужно только $O(mdq)$ шагов, что лучше, чем $O(md^2)$ для $d \gg q$
- Мы можем выполнить EM поэтапно
- Вероятностный PCA может справиться с пропущенными значениями путем их исключения по распределению по ненаблюдаемым переменным

EM против MLE

- Преимущество итеративного алгоритма EM для PCA: вычислительная эффективность для крупномасштабных приложений
- PCA: $O(d^3)$ для собственного разложения или $O(qd^2)$, если нам нужны первые q собственных векторов
- Однако, нам нужно $O(md^2)$ для вычисления ковариационной матрицы.
- В случае алгоритма EM нам нужно только $O(mdq)$ шагов, что лучше, чем $O(md^2)$ для $d \gg q$
- Мы можем выполнить EM поэтапно
- Вероятностный PCA может справиться с пропущенными значениями путем их исключения по распределению по ненаблюдаемым переменным

EM против MLE

- Преимущество итеративного алгоритма EM для PCA: вычислительная эффективность для крупномасштабных приложений
- PCA: $O(d^3)$ для собственного разложения или $O(qd^2)$, если нам нужны первые q собственных векторов
- Однако, нам нужно $O(md^2)$ для вычисления ковариационной матрицы.
- В случае алгоритма EM нам нужно только $O(mdq)$ шагов, что лучше, чем $O(md^2)$ для $d \gg q$
- Мы можем выполнить EM поэтапно
- Вероятностный PCA может справиться с пропущенными значениями путем их исключения по распределению по ненаблюдаемым переменным

EM против MLE

- Преимущество итеративного алгоритма EM для PCA: вычислительная эффективность для крупномасштабных приложений
- PCA: $O(d^3)$ для собственного разложения или $O(qd^2)$, если нам нужны первые q собственных векторов
- Однако, нам нужно $O(md^2)$ для вычисления ковариационной матрицы.
- В случае алгоритма EM нам нужно только $O(mdq)$ шагов, что лучше, чем $O(md^2)$ для $d \gg q$
- Мы можем выполнить EM поэтапно
- Вероятностный PCA может справиться с пропущенными значениями путем их исключения по распределению по ненаблюдаемым переменным

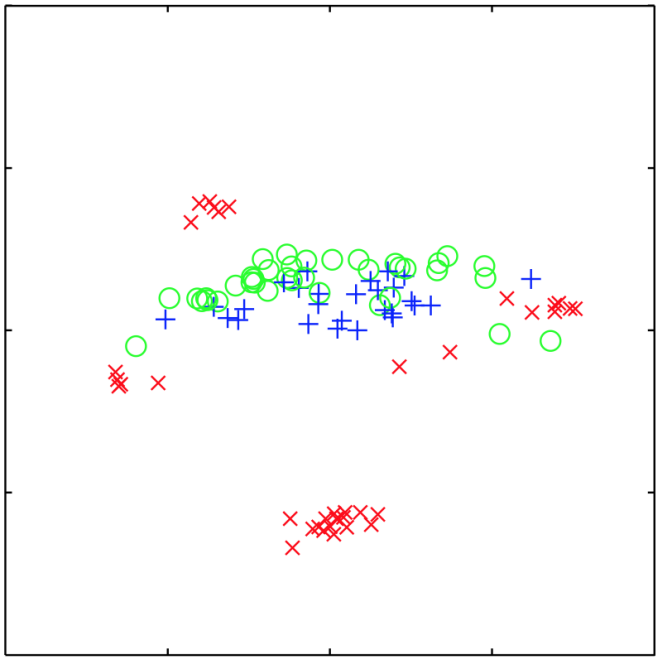
EM против MLE

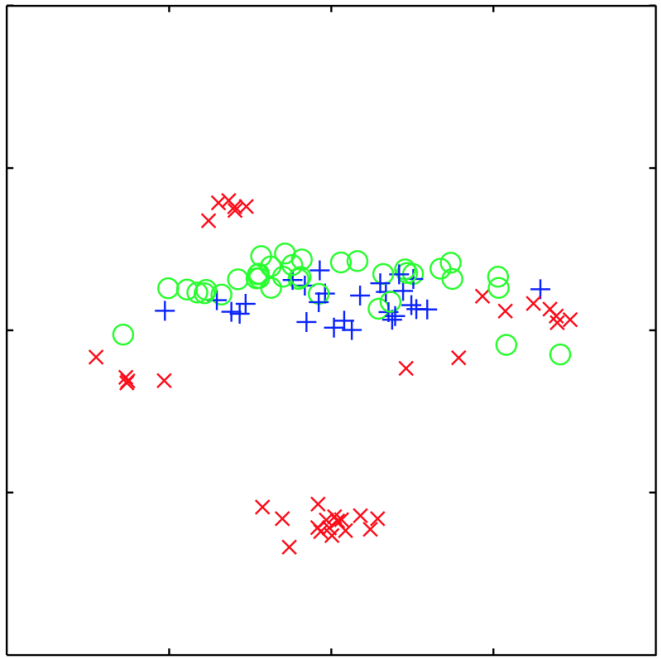
- Преимущество итеративного алгоритма EM для PCA: вычислительная эффективность для крупномасштабных приложений
- PCA: $O(d^3)$ для собственного разложения или $O(qd^2)$, если нам нужны первые q собственных векторов
- Однако, нам нужно $O(md^2)$ для вычисления ковариационной матрицы.
- В случае алгоритма EM нам нужно только $O(mdq)$ шагов, что лучше, чем $O(md^2)$ для $d \gg q$
- Мы можем выполнить EM поэтапно
- Вероятностный PCA может справиться с пропущенными значениями путем их исключения по распределению по ненаблюдаемым переменным

EM против MLE

- Преимущество итеративного алгоритма EM для PCA: вычислительная эффективность для крупномасштабных приложений
- PCA: $O(d^3)$ для собственного разложения или $O(qd^2)$, если нам нужны первые q собственных векторов
- Однако, нам нужно $O(md^2)$ для вычисления ковариационной матрицы.
- В случае алгоритма EM нам нужно только $O(mdq)$ шагов, что лучше, чем $O(md^2)$ для $d \gg q$
- Мы можем выполнить EM поэтапно
- Вероятностный PCA может справиться с пропущенными значениями путем их исключения по распределению по ненаблюдаемым переменным

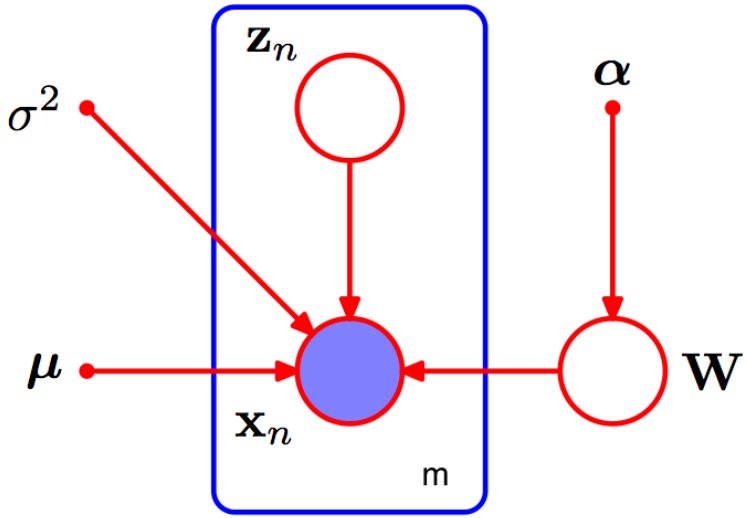
Эффективное количество параметров





- Вероятностный PCA: визуализация 100 точек данных.
- Слева: апостприорные средние проекции точек данных на главное подпространство.
- Справа: получилось путем случайного пропуска 30% значений переменной и последующего использования EM для обработки пропущенных значений

Выбор модели PCA



- Как выбрать q?
- Нам необходимо маргинализировать параметры модели $\boldsymbol{\mu}$, $\mathbf{W}$ и σ^2
- Здесь мы рассмотрим более простой подход: аппроксимации обоснованности
- $\boldsymbol{\alpha}$ определяет, какие скрытые измерения следует сократить

Выбор модели PCA

- Мы используем ARD (автоматическое определение релевантности), которое позволяет исключать из модели избыточные измерения в основном подпространстве.

$$p(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{i=1}^{q} \left(\frac{\alpha_i}{2\pi}\right)^{d/2} \exp\left\{-\frac{1}{2}\alpha_i \mathbf{w}_i^\top \mathbf{w}_i\right\}$$

- Значения α_i переоцениваются во время обучения путем максимизации логарифмической предельной вероятности, определяемой

$$p(\mathbf{X}_m|\boldsymbol{\alpha}, \boldsymbol{\mu}, \sigma^2) = \int p(\mathbf{X}_m|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) p(\mathbf{W}|\boldsymbol{\alpha}) d\mathbf{W}$$

Выбор модели PCA

Поскольку интеграл не отслеживается, мы используем аппроксимацию Лапласа и алгоритм итеративной оценки:

- Инициализировать α_i
- Применить EM-алгоритм для оценки $\mathbf{W}$ и σ^2. Единственное изменение в M-шаговое уравнение для $\mathbf{W}$

$$\mathbf{W}_{\text{new}} = \left[\sum_{n=1}^{m}(\mathbf{x}_n - \overline{\mathbf{x}})\mathbb{E}[\mathbf{z}_n]^\top\right]\left[\sum_{n=1}^{m}\mathbb{E}[\mathbf{z}_n\mathbf{z}_n^\top] + \sigma^2\boldsymbol{\alpha}\right]^{-1},$$

  где $\boldsymbol{\alpha} = \text{diag}(\alpha_i)$. Значение $\boldsymbol{\mu}$ определяется средним значением выборки, как и раньше
- Пересчитать α_i максимизируя $p(\mathbf{X}_m|\boldsymbol{\alpha}, \boldsymbol{\mu}, \sigma^2)$:

$$\alpha_i^{\text{new}} = \frac{d}{\mathbf{w}_i^\top \mathbf{w}_i}$$

- Обычно мы начинаем с некоторого $q \leq d-1$. Если некоторые α_i стремятся к бесконечности, мы можем удалить соответствующие измерения

Выводы

EM может

- заполнить недостающие данные
- выявить структуру данных (многообразия, кластеры)
- найти скрытую информацию в наборе данных для обучения
- обрабатывать неизвестные факторы, вызванные нашим выбором θ, например, в методах обучения с подкреплением
- использоваться для построения более гибких моделей данных с лучшими способностями прогнозирования
- использоваться для больших наборов данных, так как время обучения примерно такое же, как и для аналогичных моделей без скрытых переменных

Выводы

EM может

- заполнить недостающие данные
- выявить структуру данных (многообразия, кластеры)
- найти скрытую информацию в наборе данных для обучения
- обрабатывать неизвестные факторы, вызванные нашим выбором θ, например, в методах обучения с подкреплением
- использоваться для построения более гибких моделей данных с лучшими способностями прогнозирования
- использоваться для больших наборов данных, так как время обучения примерно такое же, как и для аналогичных моделей без скрытых переменных

Выводы

EM может

- заполнить недостающие данные
- выявить структуру данных (многообразия, кластеры)
- найти скрытую информацию в наборе данных для обучения
- обрабатывать неизвестные факторы, вызванные нашим выбором θ, например, в методах обучения с подкреплением
- использоваться для построения более гибких моделей данных с лучшими способностями прогнозирования
- использоваться для больших наборов данных, так как время обучения примерно такое же, как и для аналогичных моделей без скрытых переменных

Выводы

EM может

- заполнить недостающие данные
- выявить структуру данных (многообразия, кластеры)
- найти скрытую информацию в наборе данных для обучения
- обрабатывать неизвестные факторы, вызванные нашим выбором $\boldsymbol{\theta}$, например, в методах обучения с подкреплением
- использоваться для построения более гибких моделей данных с лучшими способностями прогнозирования
- использоваться для больших наборов данных, так как время обучения примерно такое же, как и для аналогичных моделей без скрытых переменных

Выводы

EM может

- заполнить недостающие данные
- выявить структуру данных (многообразия, кластеры)
- найти скрытую информацию в наборе данных для обучения
- обрабатывать неизвестные факторы, вызванные нашим выбором $\boldsymbol{\theta}$, например, в методах обучения с подкреплением
- использоваться для построения более гибких моделей данных с лучшими способностями прогнозирования
- использоваться для больших наборов данных, так как время обучения примерно такое же, как и для аналогичных моделей без скрытых переменных

Выводы

EM может

- заполнить недостающие данные
- выявить структуру данных (многообразия, кластеры)
- найти скрытую информацию в наборе данных для обучения
- обрабатывать неизвестные факторы, вызванные нашим выбором $\boldsymbol{\theta}$, например, в методах обучения с подкреплением
- использоваться для построения более гибких моделей данных с лучшими способностями прогнозирования
- использоваться для больших наборов данных, так как время обучения примерно такое же, как и для аналогичных моделей без скрытых переменных