Script **generated by TTT**

Title: groh: profile1 (02.06.2015)

Date: Tue Jun 02 15:20:48 CEST 2015

Duration: 69:52 min

Pages: 61

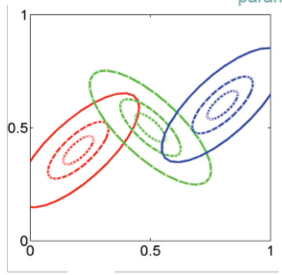
Gaussian Mixture Models

- Fuzzy C-Means is “OK” as a non-crisp clustering alg. but (as K-Means) favors spherical clusters → better approaches

- Example: Gaussian Mixture Models (GMM)
  - Linear combination of Gaussians

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \quad \text{where} \quad \sum_{k=1}^{K} \pi_k = 1, \quad 0 \leqslant \pi_k \leqslant 1$$

parameters to be estimated



[6]

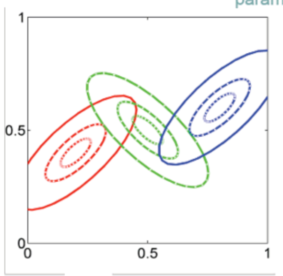
Gaussian Mixture Models

- Fuzzy C-Means is “OK” as a non-crisp clustering alg. but (as K-Means) favors spherical clusters → better approaches

- Example: Gaussian Mixture Models (GMM)
  - Linear combination of Gaussians

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \quad \text{where} \quad \sum_{k=1}^{K} \pi_k = 1, \quad 0 \leqslant \pi_k \leqslant 1$$

parameters to be estimated



[6]

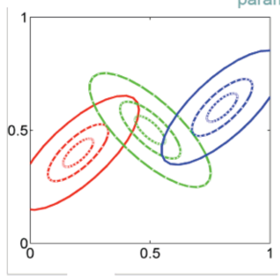
Gaussian Mixture Models

- Fuzzy C-Means is “OK” as a non-crisp clustering alg. but (as K-Means) favors spherical clusters → better approaches

- Example: Gaussian Mixture Models (GMM)
  - Linear combination of Gaussians

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \quad \text{where} \quad \sum_{k=1}^{K} \pi_k = 1, \quad 0 \leqslant \pi_k \leqslant 1$$

parameters to be estimated



[6]

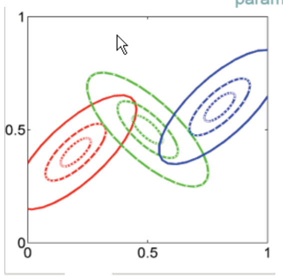
Gaussian Mixture Models

- Fuzzy C-Means is “OK” as a non-crisp clustering alg. but (as K-Means) favors spherical clusters → better approaches

- Example: Gaussian Mixture Models (GMM)
  - Linear combination of Gaussians

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \quad \text{where} \quad \sum_{k=1}^{K} \pi_k = 1, \quad 0 \leqslant \pi_k \leqslant 1$$

parameters to be estimated



[6]

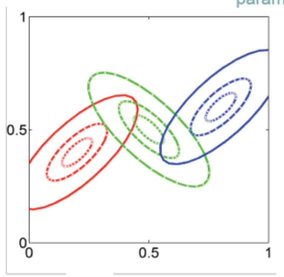
Gaussian Mixture Models

- Fuzzy C-Means is “OK” as a non-crisp clustering alg. but (as K-Means) favors spherical clusters → better approaches

- Example: Gaussian Mixture Models (GMM)
  - Linear combination of Gaussians

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \quad \text{where} \quad \sum_{k=1}^{K} \pi_k = 1, \quad 0 \leqslant \pi_k \leqslant 1$$

parameters to be estimated



[6]

Machine Learning

- For a distribution $p(x|\Theta)$ parametrized by a set of parameters Θ and iid data $X = \{x_1, x_2, \dots, x_N\}$, simple machine learning corresponds to finding the Θ that best explains the data

- iid: „identically independently drawn“ ⇨ $p(X|\Theta) = \prod_i p(x_i|\Theta)$

- $p(X|\Theta)$ is called likelihood

- „finding the Θ that best explains the data“:
  Maximum Likelihood: $\Theta_{ML} = argmax_\Theta \; p(X|\Theta)$ ⇨ $\nabla_\Theta \, p(X|\Theta) \stackrel{!}{=} 0$

- convenient: use $\log \, p(X|\theta)$ instead of $p(X|\theta)$
  ⇨ $\log \; p(X|\Theta) = \sum_i \log \, p(x\_i|\Theta)$

Machine Learning

- For a distribution $p(x|\Theta)$ parametrized by a set of parameters Θ and iid data $X = \{x_1, x_2, \dots, x_N\}$, simple machine learning corresponds to finding the Θ that best explains the data

- iid: „identically independently drawn“ ⇨ $p(X|\Theta) = \prod_i p(x_i|\Theta)$

- $p(X|\Theta)$ is called likelihood

- „finding the Θ that best explains the data“:
  Maximum Likelihood: $\Theta_{ML} = argmax_\Theta \; p(X|\Theta)$ ⇨ $\nabla_\Theta \, p(X|\Theta) \stackrel{!}{=} 0$

- convenient: use $\log \, p(X|\theta)$ instead of $p(X|\theta)$
  ⇨ $\log \; p(X|\Theta) = \sum_i \log \, p(x\_i|\Theta)$

Machine Learning

- For a distribution $p(x|\Theta)$ parametrized by a set of parameters Θ and iid data $X = \{x_1, x_2, \dots, x_N\}$, simple machine learning corresponds to finding the Θ that best explains the data
- iid: „identically independently drawn“ ⇨ $p(X|\Theta) = \prod_i p(x_i|\Theta)$
- $p(\mathrm{X}|\Theta)$ is called likelihood
- „finding the Θ that best explains the data“: Maximum Likelihood: $\Theta_{ML} = argmax_\Theta \; p(X|\Theta) \;\; ⇨ \;\; \nabla_\Theta \, p(X|\Theta) \overset{!}{=} 0$
- convenient: use $\log \; p(X|\theta)$ instead of $p(X|\theta)$
  ⇨ $\log \; p(X|\Theta) = \sum_i \log \, p(x\_i|\Theta)$

Machine Learning

- For a distribution $p(x|\Theta)$ parametrized by a set of parameters Θ and iid data $X = \{x_1, x_2, \dots, x_N\}$, simple machine learning corresponds to finding the Θ that best explains the data
- iid: „identically independently drawn“ ⇨ $p(X|\Theta) = \prod_i p(x_i|\Theta)$
- $p(\mathrm{X}|\Theta)$ is called likelihood
- „finding the Θ that best explains the data“: Maximum Likelihood: $\Theta_{ML} = argmax_\Theta \; p(X|\Theta) \;\; ⇨ \;\; \nabla_\Theta \, p(X|\Theta) \overset{!}{=} 0$
- convenient: use $\log \; p(X|\theta)$ instead of $p(X|\theta)$
  ⇨ $\log \; p(X|\Theta) = \sum_i \log \, p(x\_i|\Theta)$

Machine Learning

- For a distribution $p(x|\Theta)$ parametrized by a set of parameters Θ and iid data $X = \{x_1, x_2, \dots, x_N\}$, simple machine learning corresponds to finding the Θ that best explains the data
- iid: „identically independently drawn“ ⇨ $p(X|\Theta) = \prod_i p(x_i|\Theta)$
- $p(\mathrm{X}|\Theta)$ is called likelihood
- „finding the Θ that best explains the data“: Maximum Likelihood: $\Theta_{ML} = argmax_\Theta \; p(X|\Theta) \;\; ⇨ \;\; \nabla_\Theta \, p(X|\Theta) \overset{!}{=} 0$
- convenient: use $\log \; p(X|\theta)$ instead of $p(X|\theta)$
  ⇨ $\log \; p(X|\Theta) = \sum_i \log \, p(x\_i|\Theta)$

Machine Learning

- For a distribution $p(x|\Theta)$ parametrized by a set of parameters Θ and iid data $X = \{x_1, x_2, \dots, x_N\}$, simple machine learning corresponds to finding the Θ that best explains the data
- iid: „identically independently drawn“ ⇨ $p(X|\Theta) = \prod_i p(x_i|\Theta)$
- $p(\mathrm{X}|\Theta)$ is called likelihood
- „finding the Θ that best explains the data“: Maximum Likelihood: $\Theta_{ML} = argmax_\Theta \; p(X|\Theta) \;\; ⇨ \;\; \nabla_\Theta \, p(X|\Theta) \overset{!}{=} 0$
- convenient: use $\log \; p(X|\theta)$ instead of $p(X|\theta)$
  ⇨ $\log \; p(X|\Theta) = \sum_i \log \, p(x\_i|\Theta)$

Machine Learning

Example: $x \in \mathbb{R}^m$ and $p(x|\Theta)$ is one multivariate Gaussian)

$$p(\mathbf{x}|\Theta) = \mathcal{N}(\mathbf{x}|\underbrace{\boldsymbol{\mu}, \boldsymbol{\Sigma}}_{\theta}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- log likelihood: (use base e)

$$\ln\, p(\mathbf{X}|\Theta) = \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

- Maximum log likelihood:

$$\Theta_{ML} = argmax_{\Theta}\ \log p(X|\Theta) \Rightarrow \nabla_{\Theta}\left(\textstyle\sum_i \log p(x_i|\Theta)\right) \stackrel{!}{=} 0$$

$$\left.\begin{aligned} \boldsymbol{\mu}_{\mathrm{ML}} &: \frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}\,|\,\mu, \boldsymbol{\Sigma}) = 0 \\ \boldsymbol{\Sigma}_{\mathrm{ML}} &: \frac{\partial}{\partial \boldsymbol{\Sigma}} \ln p(\mathbf{X}\,|\,\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0 \end{aligned}\right\} \Rightarrow \begin{aligned} \boldsymbol{\mu}_{\mathrm{ML}} &= \frac{1}{N}\sum^{N}\mathbf{x}_n \\ \boldsymbol{\Sigma}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}} \end{aligned}$$

Machine Learning

Example: $x \in \mathbb{R}^m$ and $p(x|\Theta)$ is one multivariate Gaussian)

$$p(\mathbf{x}|\Theta) = \mathcal{N}(\mathbf{x}|\underbrace{\boldsymbol{\mu}, \boldsymbol{\Sigma}}_{\theta}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- log likelihood: (use base e)

$$\ln\, p(\mathbf{X}|\Theta) = \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

- Maximum log likelihood:

$$\Theta_{ML} = argmax_{\Theta}\ \log p(X|\Theta) \Rightarrow \nabla_{\Theta}\left(\textstyle\sum_i \log p(x_i|\Theta)\right) \stackrel{!}{=} 0$$

$$\left.\begin{aligned} \boldsymbol{\mu}_{\mathrm{ML}} &: \frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}\,|\,\mu, \boldsymbol{\Sigma}) = 0 \\ \boldsymbol{\Sigma}_{\mathrm{ML}} &: \frac{\partial}{\partial \boldsymbol{\Sigma}} \ln p(\mathbf{X}\,|\,\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0 \end{aligned}\right\} \Rightarrow \begin{aligned} \boldsymbol{\mu}_{\mathrm{ML}} &= \frac{1}{N}\sum^{N}\mathbf{x}_n \\ \boldsymbol{\Sigma}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}} \end{aligned}$$

Machine Learning

Example: $x \in \mathbb{R}^m$ and $p(x|\Theta)$ is one multivariate Gaussian)

$$p(\mathbf{x}|\Theta) = \mathcal{N}(\mathbf{x}|\underbrace{\boldsymbol{\mu}, \boldsymbol{\Sigma}}_{\theta}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- log likelihood: (use base e)

$$\ln\, p(\mathbf{X}|\Theta) = \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

- Maximum log likelihood:

$$\Theta_{ML} = argmax_{\Theta}\ \log p(X|\Theta) \Rightarrow \nabla_{\Theta}\left(\textstyle\sum_i \log p(x_i|\Theta)\right) \stackrel{!}{=} 0$$

$$\left.\begin{aligned} \boldsymbol{\mu}_{\mathrm{ML}} &: \frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}\,|\,\mu, \boldsymbol{\Sigma}) = 0 \\ \boldsymbol{\Sigma}_{\mathrm{ML}} &: \frac{\partial}{\partial \boldsymbol{\Sigma}} \ln p(\mathbf{X}\,|\,\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0 \end{aligned}\right\} \Rightarrow \begin{aligned} \boldsymbol{\mu}_{\mathrm{ML}} &= \frac{1}{N}\sum^{N}\mathbf{x}_n \\ \boldsymbol{\Sigma}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}} \end{aligned}$$

Machine Learning

Example: $x \in \mathbb{R}^m$ and $p(x|\Theta)$ is one multivariate Gaussian)

$$p(\mathbf{x}|\Theta) = \mathcal{N}(\mathbf{x}|\underbrace{\boldsymbol{\mu}, \boldsymbol{\Sigma}}_{\theta}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- log likelihood: (use base e)

$$\ln\, p(\mathbf{X}|\Theta) = \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

- Maximum log likelihood:

$$\Theta_{ML} = argmax_{\Theta}\ \log p(X|\Theta) \Rightarrow \nabla_{\Theta}\left(\textstyle\sum_i \log p(x_i|\Theta)\right) \stackrel{!}{=} 0$$

$$\left.\begin{aligned} \boldsymbol{\mu}_{\mathrm{ML}} &: \frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}\,|\,\mu, \boldsymbol{\Sigma}) = 0 \\ \boldsymbol{\Sigma}_{\mathrm{ML}} &: \frac{\partial}{\partial \boldsymbol{\Sigma}} \ln p(\mathbf{X}\,|\,\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0 \end{aligned}\right\} \Rightarrow \begin{aligned} \boldsymbol{\mu}_{\mathrm{ML}} &= \frac{1}{N}\sum^{N}\mathbf{x}_n \\ \boldsymbol{\Sigma}_{\mathrm{ML}} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}} \end{aligned}$$

Machine Learning

Example: $x \in \mathbb{R}^m$ and $p(x|\Theta)$ is one multivariate Gaussian)

$$p(\mathbf{x}|\Theta) = \mathcal{N}(\mathbf{x}|\underbrace{\boldsymbol{\mu},\boldsymbol{\Sigma}}_{\theta}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- log likelihood: (use base e)

$$\ln p(\mathbf{X}|\Theta) = \ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})$$

- Maximum log likelihood:

$$\Theta_{ML} = argmax_{\Theta}\ \log p(X|\Theta) \Rightarrow \nabla_{\Theta}\left(\sum_i \log p(x_i|\Theta)\right) \overset{!}{=} 0$$

$$\left.\begin{array}{l} \boldsymbol{\mu}_{\mathrm{ML}}:\ \dfrac{\partial}{\partial\boldsymbol{\mu}}\ln p(\mathbf{X}\,|\,\mu,\boldsymbol{\Sigma}) = 0 \\[2ex] \boldsymbol{\Sigma}_{\mathrm{ML}}:\ \dfrac{\partial}{\partial\boldsymbol{\Sigma}}\ln p(\mathbf{X}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = 0 \end{array}\right\} \Rightarrow \begin{array}{l} \boldsymbol{\mu}_{\mathrm{ML}} = \dfrac{1}{N}\displaystyle\sum^{N}\mathbf{x}_n \\[2ex] \boldsymbol{\Sigma}_{\mathrm{ML}} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}} \end{array}$$

Machine Learning

Example: $x \in \mathbb{R}^m$ and $p(x|\Theta)$ is one multivariate Gaussian)

$$p(\mathbf{x}|\Theta) = \mathcal{N}(\mathbf{x}|\underbrace{\boldsymbol{\mu},\boldsymbol{\Sigma}}_{\theta}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- log likelihood: (use base e)

$$\ln p(\mathbf{X}|\Theta) = \ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})$$

- Maximum log likelihood:

$$\Theta_{ML} = argmax_{\Theta}\ \log p(X|\Theta) \Rightarrow \nabla_{\Theta}\left(\sum_i \log p(x_i|\Theta)\right) \overset{!}{=} 0$$

$$\left.\begin{array}{l} \boldsymbol{\mu}_{\mathrm{ML}}:\ \dfrac{\partial}{\partial\boldsymbol{\mu}}\ln p(\mathbf{X}\,|\,\mu,\boldsymbol{\Sigma}) = 0 \\[2ex] \boldsymbol{\Sigma}_{\mathrm{ML}}:\ \dfrac{\partial}{\partial\boldsymbol{\Sigma}}\ln p(\mathbf{X}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = 0 \end{array}\right\} \Rightarrow \begin{array}{l} \boldsymbol{\mu}_{\mathrm{ML}} = \dfrac{1}{N}\displaystyle\sum^{N}\mathbf{x}_n \\[2ex] \boldsymbol{\Sigma}_{\mathrm{ML}} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}} \end{array}$$

Machine Learning

Example: $x \in \mathbb{R}^m$ and $p(x|\Theta)$ is one multivariate Gaussian)

$$p(\mathbf{x}|\Theta) = \mathcal{N}(\mathbf{x}|\underbrace{\boldsymbol{\mu},\boldsymbol{\Sigma}}_{\theta}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- log likelihood: (use base e)

$$\ln p(\mathbf{X}|\Theta) = \ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n-\boldsymbol{\mu})$$

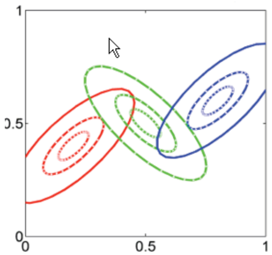
- Maximum log likelihood:

$$\Theta_{ML} = argmax_{\Theta}\ \log p(X|\Theta) \Rightarrow \nabla_{\Theta}\left(\sum_i \log p(x_i|\Theta)\right) \overset{!}{=} 0$$

$$\left.\begin{array}{l} \boldsymbol{\mu}_{\mathrm{ML}}:\ \dfrac{\partial}{\partial\boldsymbol{\mu}}\ln p(\mathbf{X}\,|\,\mu,\boldsymbol{\Sigma}) = 0 \\[2ex] \boldsymbol{\Sigma}_{\mathrm{ML}}:\ \dfrac{\partial}{\partial\boldsymbol{\Sigma}}\ln p(\mathbf{X}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = 0 \end{array}\right\} \Rightarrow \begin{array}{l} \boldsymbol{\mu}_{\mathrm{ML}} = \dfrac{1}{N}\displaystyle\sum^{N}\mathbf{x}_n \\[2ex] \boldsymbol{\Sigma}_{\mathrm{ML}} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n-\boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}} \end{array}$$

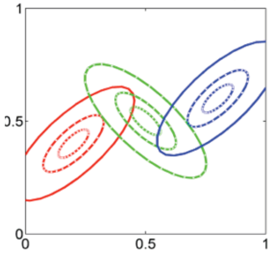
GMM-Basics

GMM: $$p(\mathbf{x}|\theta) = p(\mathbf{x}) = \sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k). \qquad 0 \leqslant \pi_k \leqslant 1 \qquad \sum_{k=1}^{K}\pi_k = 1$$



GMM-Basics

GMM: $p(\mathbf{x}|\theta) = p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \quad 0 \leqslant \pi_k \leqslant 1 \quad \sum_{k=1}^{K} \pi_k = 1$



GMM-Basics

GMM: $p(\mathbf{x}|\theta) = p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \quad 0 \leqslant \pi_k \leqslant 1 \quad \sum_{k=1}^{K} \pi_k = 1$

- **1 of K representation**

  K-dimensional binary random variable $\mathbf{z}$

  $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$

  $p(z_k = 1) = \pi_k$

  $$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

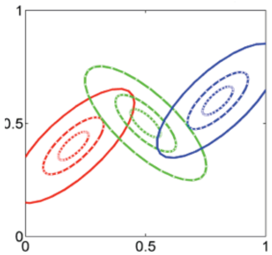
- **conditional probability**

  $p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$

  $$p(\mathbf{x}) = \sum_{\mathbf{z}} \underbrace{p(\mathbf{z})p(\mathbf{x}|\mathbf{z})}_{p(x,z)} = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

GMM-Basics

GMM: $p(\mathbf{x}|\theta) = p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \quad 0 \leqslant \pi_k \leqslant 1 \quad \sum_{k=1}^{K} \pi_k = 1$



GMM-Basics

GMM: $p(\mathbf{x}|\theta) = p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \quad 0 \leqslant \pi_k \leqslant 1 \quad \sum_{k=1}^{K} \pi_k = 1$

- **1 of K representation**

  K-dimensional binary random variable $\mathbf{z}$

  $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$

  $p(z_k = 1) = \pi_k$

  $$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

- **conditional probability**

  $p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$

  $$p(\mathbf{x}) = \sum_{\mathbf{z}} \underbrace{p(\mathbf{z})p(\mathbf{x}|\mathbf{z})}_{p(x,z)} = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

GMM-Basics

GMM: $p(\mathbf{x}|\theta) = p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \qquad 0 \leqslant \pi_k \leqslant 1 \qquad \sum_{k=1}^{K} \pi_k = 1$

- 1 of K representation

  K-dimensional binary random variable $\mathbf{z}$

  $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$

  $p(z_k = 1) = \pi_k$

  $$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

- conditional probability

  $$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

  $$p(\mathbf{x}) = \sum_{\mathbf{z}} \underbrace{p(\mathbf{z})p(\mathbf{x}|\mathbf{z})}_{p(x,z)} = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

GMM-Basics

GMM: $p(\mathbf{x}|\theta) = p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \qquad 0 \leqslant \pi_k \leqslant 1 \qquad \sum_{k=1}^{K} \pi_k = 1$

- 1 of K representation

  K-dimensional binary random variable $\mathbf{z}$

  $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$

  $p(z_k = 1) = \pi_k$

  $$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

- conditional probability

  $$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

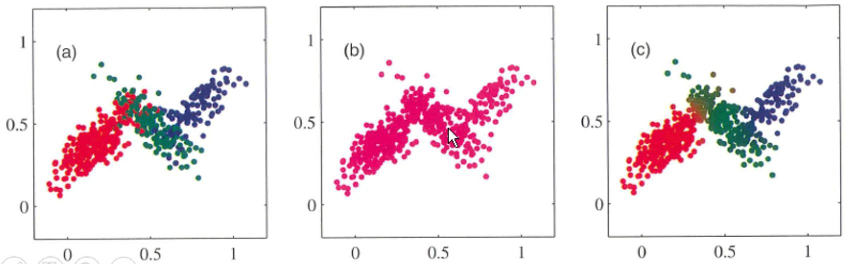
  $$p(\mathbf{x}) = \sum_{\mathbf{z}} \underbrace{p(\mathbf{z})p(\mathbf{x}|\mathbf{z})}_{p(x,z)} = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

GMM-Basics

- Responsibilities

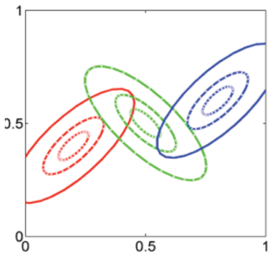
$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

- Example



GMM-Basics

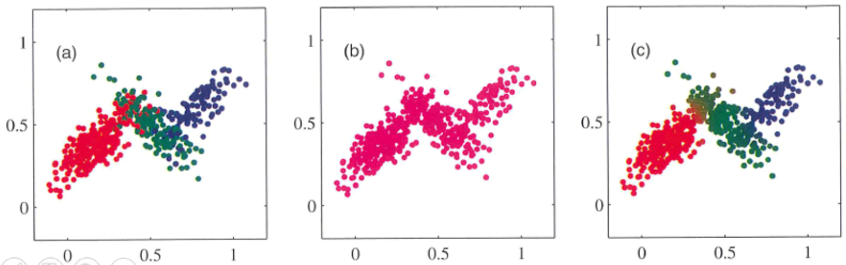
GMM: $p(\mathbf{x}|\theta) = p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \qquad 0 \leqslant \pi_k \leqslant 1 \qquad \sum_{k=1}^{K} \pi_k = 1$



- **Responsibilities**

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

- **Example**

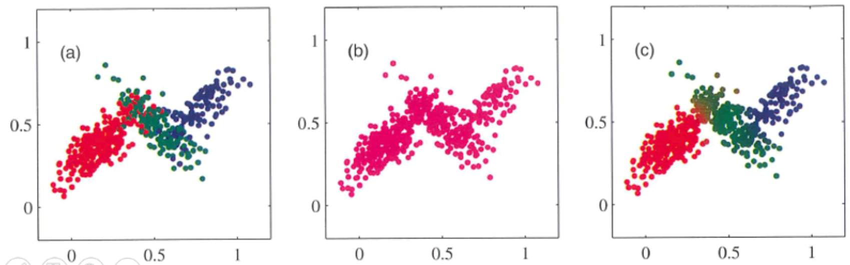


GMM-Basics

- **Responsibilities**

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

- **Example**

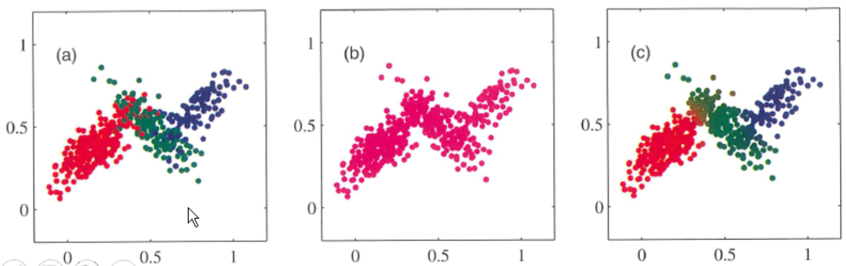


GMM-Basics

- **Responsibilities**

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

- **Example**

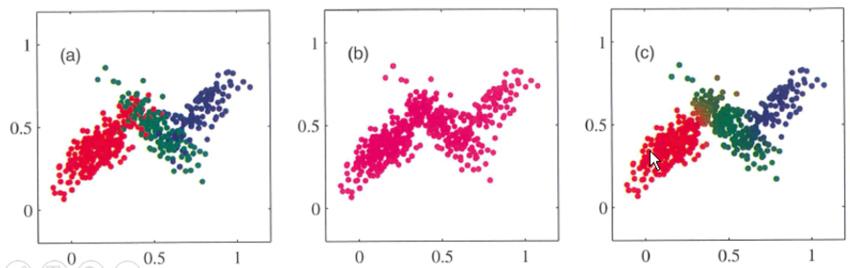


GMM-Basics

- **Responsibilities**

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

- **Example**

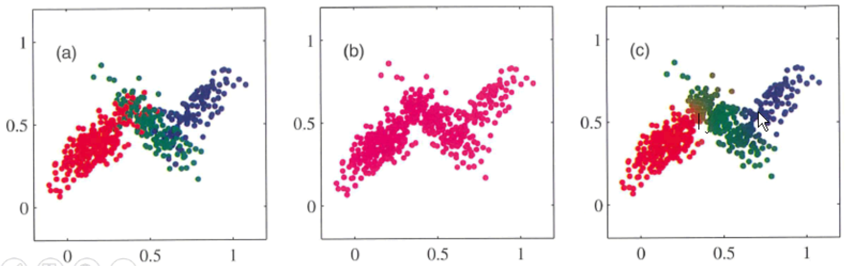


GMM-Basics

- **Responsibilities**

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

- **Example**



GMM-Basics

Maximum likelihood (GMM)

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Vector of K *D*-dim. means $\boldsymbol{\mu}_k$

Vector of K *DxD* covariances $\boldsymbol{\Sigma}_k$

- maximizing w.r.t $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ →

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \qquad \boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\left(N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \right) \qquad \pi_k = \frac{N_k}{N}$$

GMM-Basics

Maximum likelihood (GMM)

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Vector of K *D*-dim. means $\boldsymbol{\mu}_k$

Vector of K *DxD* covariances $\boldsymbol{\Sigma}_k$

- maximizing w.r.t $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ →

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \qquad \boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\left(N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \right) \qquad \pi_k = \frac{N_k}{N}$$

GMM-Basics

Maximum likelihood (GMM)

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Vector of K *D*-dim. means $\boldsymbol{\mu}_k$

Vector of K *DxD* covariances $\boldsymbol{\Sigma}_k$

- maximizing w.r.t $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ →

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \qquad \boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\left(N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \right) \qquad \pi_k = \frac{N_k}{N}$$

GMM-Basics

Maximum likelihood (GMM)

$$\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Vector of K D-dim. means $\boldsymbol{\mu}_\mathrm{k}$

Vector of K DxD covariances $\boldsymbol{\Sigma}_\mathrm{k}$

- maximizing w.r.t $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ →

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \qquad \boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\left(N_k = \sum_{n=1}^{N}\gamma(z_{nk}). \right) \qquad \pi_k = \frac{N_k}{N}$$

GMM-Basics

Maximum likelihood (GMM)

$$\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Vector of K D-dim. means $\boldsymbol{\mu}_\mathrm{k}$

Vector of K DxD covariances $\boldsymbol{\Sigma}_\mathrm{k}$

- maximizing w.r.t $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ →

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \qquad \boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\left(N_k = \sum_{n=1}^{N}\gamma(z_{nk}). \right) \qquad \pi_k = \frac{N_k}{N}$$

GMM-Basics

Maximum likelihood (GMM)

$$\ln p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Vector of K D-dim. means $\boldsymbol{\mu}_\mathrm{k}$

Vector of K DxD covariances $\boldsymbol{\Sigma}_\mathrm{k}$

- maximizing w.r.t $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ →

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \qquad \boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n-\boldsymbol{\mu}_k)(\mathbf{x}_n-\boldsymbol{\mu}_k)^{\mathrm{T}}$$

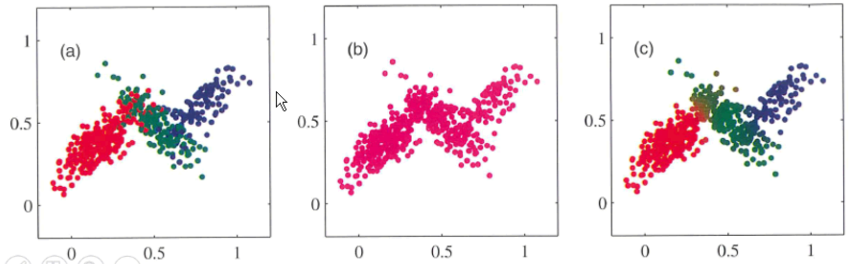
$$\left(N_k = \sum_{n=1}^{N}\gamma(z_{nk}). \right) \qquad \pi_k = \frac{N_k}{N}$$

GMM-Basics

- Responsibilities

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k=1)p(\mathbf{x}|z_k=1)}{\sum_{j=1}^{K} p(z_j=1)p(\mathbf{x}|z_j=1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}.$$

- Example



Maximum likelihood (GMM)

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Vector of K D-dim. means $\boldsymbol{\mu}_k$

Vector of K DxD covariances $\boldsymbol{\Sigma}_k$

- maximizing w.r.t $\boldsymbol{\pi}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ →

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \qquad \boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\left(N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \right) \qquad \pi_k = \frac{N_k}{N}$$

Maximum likelihood (GMM)

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \qquad \boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

$$\left(N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \right) \qquad \pi_k = \frac{N_k}{N}$$

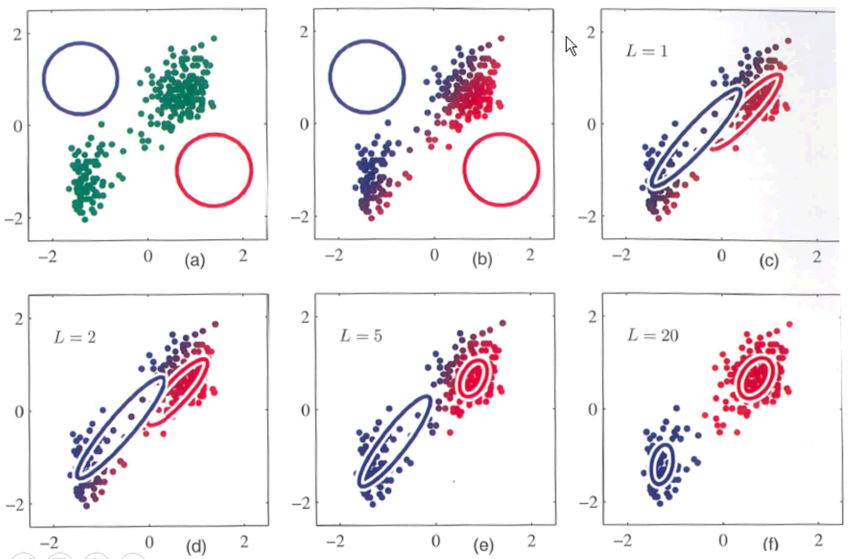
- so what?! → Problem: Expr. depend on $\gamma(z_{nk})$ which depends on $\boldsymbol{\pi}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ which depends on $\gamma(z_{nk})$ which depends on .....

- Idea: Alternating approach (EM-algorithm):

  Step t: Evaluate $\gamma(z_{nk})_{(t)}$ using $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})_{(t-1)}$

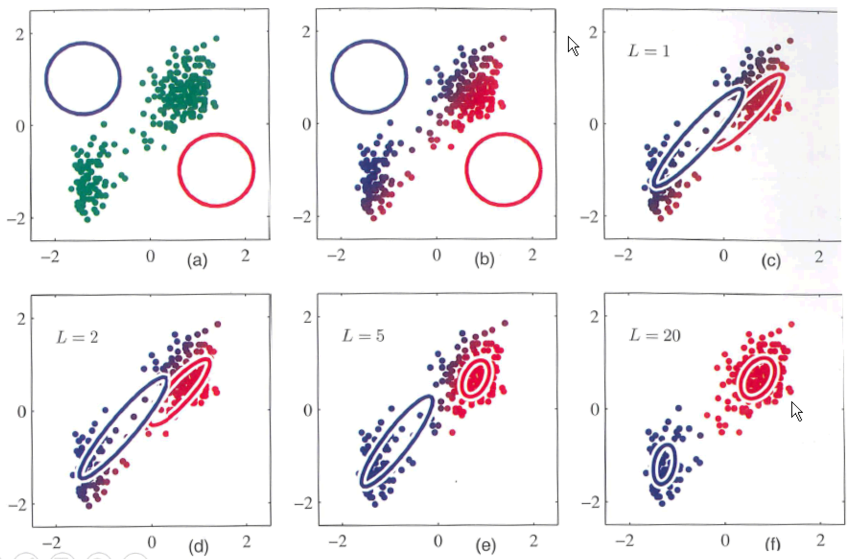
  Evaluate $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})_{(t)}$ using $\gamma(z_{nk})_{(t)}$

Maximum likelihood (GMM)



GMM-Basics

Maximum likelihood (GMM)



EM-algorithm: General View

- Having latent variables $\mathbf{Z}$, ML becomes

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\}$$

- Summation inside $\ln$ → Problems !

- If we knew the complete dataset $\{\mathbf{X}, \mathbf{Z}\}$ (and thus the distribution $p(\mathbf{X}, \mathbf{Z}|\theta)$), we could use ML to solve for θ with $p(\mathbf{X}, \mathbf{Z}|\theta)$ directly (which is easy, as we will see, because $p(\mathbf{X}, \mathbf{Z}|\theta)$ is of exponential family (the functional form is known!!)

- We only know $p(\mathbf{Z}|\mathbf{X},\theta)$ (→ responsibilities, as we will see) → compute expectation of (unknown) quantity $p(\mathbf{X}, \mathbf{Z}|\theta)$ or even better of the quantity $\ln p(\mathbf{X}, \mathbf{Z}|\theta)$

EM-algorithm: General View

- alternating EM:

E-Step: compute

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}).$$

M-Step: compute

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}).$$

EM-algorithm: General View

- alternating EM:

E-Step: compute

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}).$$

M-Step: compute

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}).$$

EM: Relation to K-Means

- If we use k Gaussians with $\Sigma = \varepsilon I$:

that is why K-Means favors spherical clusters

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{(2\pi\epsilon)^{1/2}} \exp\left\{ -\frac{1}{2\epsilon} \|\mathbf{x} - \boldsymbol{\mu}_k\|^2 \right\}$$

- we get for the responsibilities:

$$\gamma(z_{nk}) = \frac{\pi_k \exp\{-\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2/2\epsilon\}}{\sum_j \pi_j \exp\{-\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2/2\epsilon\}}$$

- Letting ε → 0 and Taylor-Expansion:

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \Sigma, \boldsymbol{\pi})] \to -\frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 + \text{const}$$

→ same as on slide 18

EM: Relation to K-Means

- If we use k Gaussians with Σ = εI:

that is why K-Means favors spherical clusters

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{1/2}} \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x}-\boldsymbol{\mu}_k\|^2\right\}$$

- we get for the responsibilities:

$$\gamma(z_{nk}) = \frac{\pi_k \exp\{-\|\mathbf{x}_n-\boldsymbol{\mu}_k\|^2/2\epsilon\}}{\sum_j \pi_j \exp\{-\|\mathbf{x}_n-\boldsymbol{\mu}_j\|^2/2\epsilon\}}$$

- Letting ε → 0 and Taylor-Expansion:

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})] \to -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n-\boldsymbol{\mu}_k\|^2 + \text{const}$$

→ same as on slide 18

EM: Relation to K-Means

- If we use k Gaussians with Σ = εI:

that is why K-Means favors spherical clusters

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{1/2}} \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x}-\boldsymbol{\mu}_k\|^2\right\}$$

- we get for the responsibilities:

$$\gamma(z_{nk}) = \frac{\pi_k \exp\{-\|\mathbf{x}_n-\boldsymbol{\mu}_k\|^2/2\epsilon\}}{\sum_j \pi_j \exp\{-\|\mathbf{x}_n-\boldsymbol{\mu}_j\|^2/2\epsilon\}}$$

- Letting ε → 0 and Taylor-Expansion:

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi})] \to -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n-\boldsymbol{\mu}_k\|^2 + \text{const}$$

→ same as on slide 18

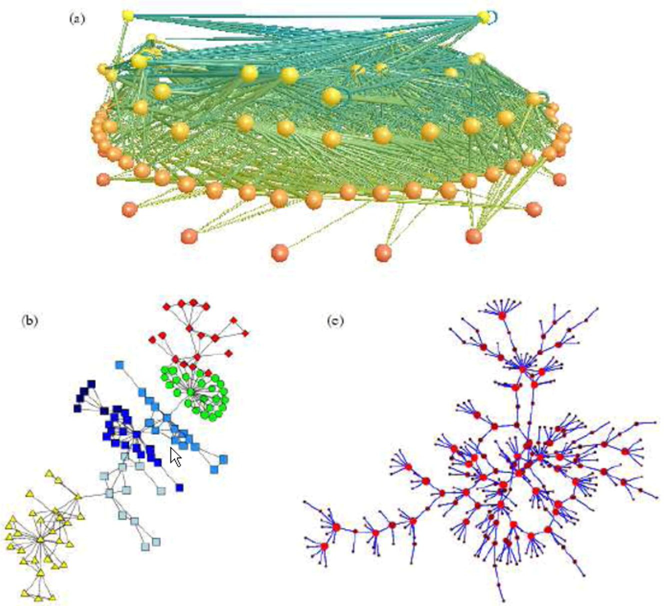


FIG. 2 Three examples of the kinds of networks that are the topic of this review. (a) A food web of predator-prey interactions between species in a freshwater lake [272]. Picture courtesy of Neo Martinez and Richard Williams. (b) The network of collaborations between scientists at a private research institution [171]. (c) A network of sexual contacts between individuals in the study by Potterat *et al.* [342].

[1]

Studying Complex Networks

- Paradigm shift: small NW → large NW:
  - interest in individual elements (“centrality of node x”) → interest in global statistical / topological properties (“degree distribution of NW”)
  - investigating particular NW instance → general model for types of NW with certain properties
  - 100 nodes → 10^8 nodes
  - visualization possible → impossible / pointless

- Typical sorts of NW investigated:
  social, information, technological, biological

Studying Complex Networks

- Paradigm shift: small NW → large NW:
  - interest in individual elements ("centrality of node x") → interest in global statistical / topological properties ("degree distribution of NW")
  - investigating particular NW instance → general model for types of NW with certain properties
  - 100 nodes → 10^8 nodes
  - visualization possible → impossible / pointless

- Typical sorts of NW investigated:
  social, information, technological, biological

Social Networks

- Formalizations of social context (mostly long and medium term)

- Until dawn of databases : Collection via social science methods (questionnaires etc.) → „laboratory effects" ;
- Until dawn of Social Web: „Indirect characterization of social relations" (e.g. co-citation networks, collaboration NW (e.g. movie co-acting NW))
- Today (Social Web / Mobile Social Web) : self declared explicification of social structures ; „collect data about" / „observe" Homo Sapiens in its „natural habitat" (→ Twitter, Facebook etc.)

Information Networks / Knowledge NW

- Most studied examples: citation NW (tree), the WWW;
- Example findings:
  - p(k) of author having k papers: $p(k) \sim k^{-\alpha}$: power law
  - distribution of in or out degrees of WWW pages (also for citation NW): $p(k) \sim k^{-\alpha}$

- Other examples:
  - bipartite preference networks :
    → recommender systems == link prediction on these NW; example: collaborative filtering
  - ontologies, semantic NW
  - word networks
  - tripartite tag/author/item networks
    → Folksonomies

Information Networks / Knowledge NW

- Most studied examples: citation NW (tree), the WWW;
- Example findings:
  - p(k) of author having k papers: $p(k) \sim k^{-\alpha}$: power law
  - distribution of in or out degrees of WWW pages (also for citation NW): $p(k) \sim k^{-\alpha}$

- Other examples:
  - bipartite preference networks :
    → recommender systems == link prediction on these NW; example: collaborative filtering
  - ontologies, semantic NW
  - word networks
  - tripartite tag/author/item networks
    → Folksonomies

Information Networks / Knowledge NW

- Most studied examples: citation NW (tree), the WWW;
- Example findings:
  - p(k) of author having k papers: $p(k) \sim k^{-\alpha}$: power law
  - distribution of in or out degrees of WWW pages (also for citation NW): $p(k) \sim k^{-\alpha}$

- Other examples:
  - bipartite preference networks :
    → recommender systems == link prediction on these NW;
    example: collaborative filtering
  - ontologies, semantic NW
  - word networks
  - tripartite tag/author/item networks
    → Folksonomies

Technological Networks

- Most studied examples: distribution NW:
  - the Internet,
  - electric power grids,
  - traffic NW (roads, railway tracks etc.))

Biological / Chemistry / Physics Networks

- Most studied examples:
  - biochemical pathways, gene-protein and protein-protein interaction NW
  - nervous systems, vascular systems (also natural distribution NW),
  - food NW, ecological dependency NW

Technological Networks

- Most studied examples: distribution NW:
  - the Internet,
  - electric power grids,
  - traffic NW (roads, railway tracks etc.))

Biological / Chemistry / Physics Networks

- Most studied examples:
  - biochemical pathways, gene-protein and protein-protein interaction NW
  - nervous systems, vascular systems (also natural distribution NW),
  - food NW, ecological dependency NW

Technological Networks

- Most studied examples: distribution NW:
  - the Internet,
  - electric power grids,
  - traffic NW (roads, railway tracks etc.))

Biological / Chemistry / Physics Networks

- Most studied examples:
  - biochemical pathways, gene-protein and protein-protein interaction NW
  - nervous systems, vascular systems (also natural distribution NW),
  - food NW, ecological dependency NW

Technological Networks

- Most studied examples: distribution NW:
  - the Internet,
  - electric power grids,
  - traffic NW (roads, railway tracks etc.))

Biological / Chemistry / Physics Networks

- Most studied examples:
  - biochemical pathways, gene-protein and protein-protein interaction NW
  - nervous systems, vascular systems (also natural distribution NW),
  - food NW, ecological dependency NW

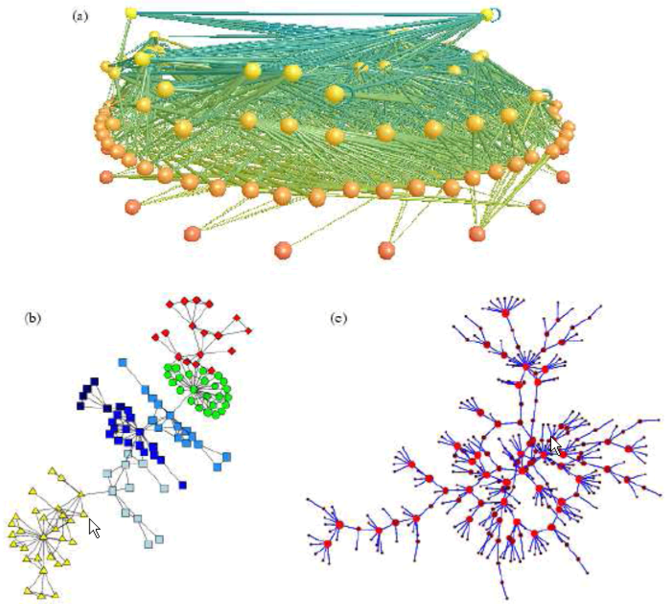


FIG. 2 Three examples of the kinds of networks that are the topic of this review. (a) A food web of predator-prey interactions between species in a freshwater lake [272]. Picture courtesy of Neo Martinez and Richard Williams. (b) The network of collaborations between scientists at a private research institution [171]. (c) A network of sexual contacts between individuals in the study by Potterat *et al.* [342].

[1]