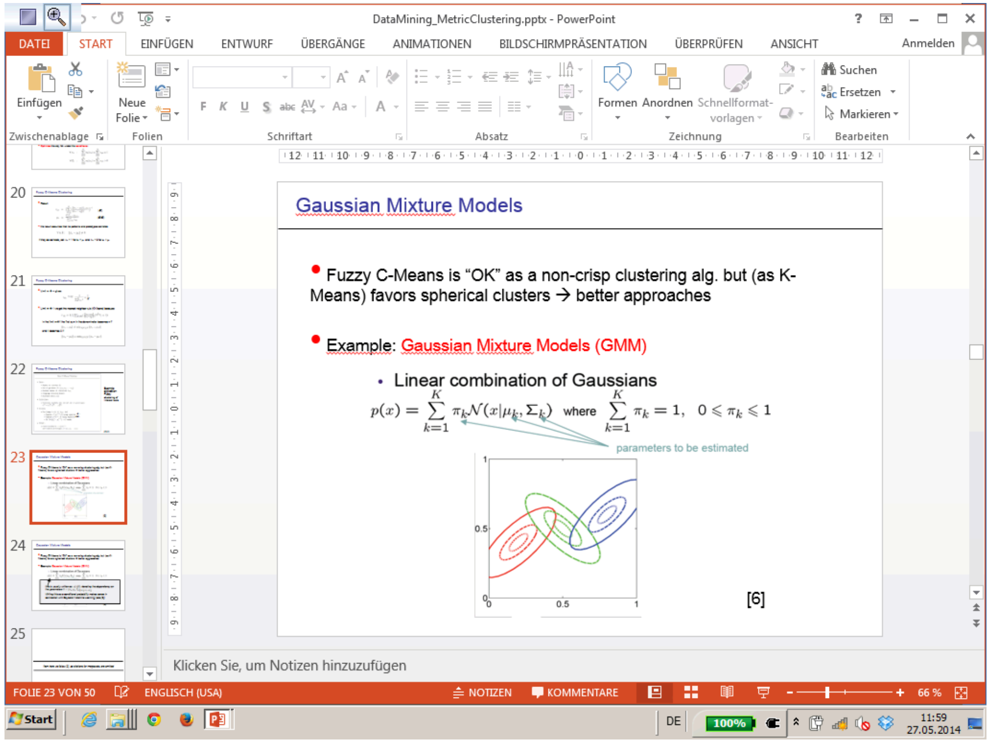
**Script** generated by TTT

Title: groh: profile1 (27.05.2014)

Date: Tue May 27 11:59:47 CEST 2014

Duration: 89:52 min

Pages: 91

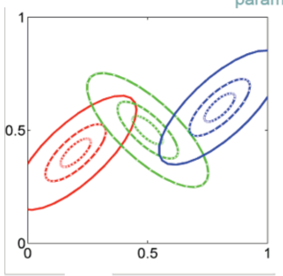



## Gaussian Mixture Models

- 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



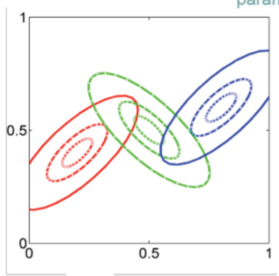

[6]


## 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

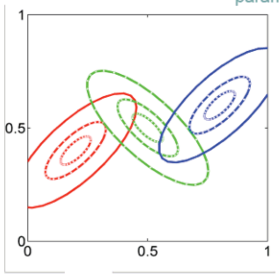
[6]

- 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]

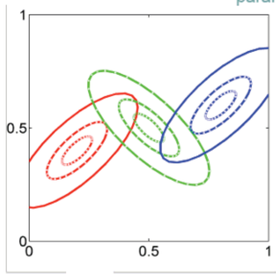
Gaussian Mixture Models

- 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]

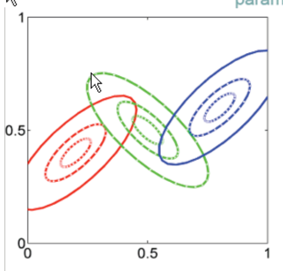
Gaussian Mixture Models

- 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]

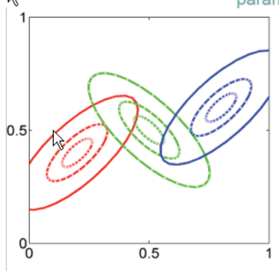
Gaussian Mixture Models

- 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]

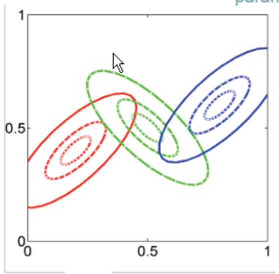
## Gaussian Mixture Models

- 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]

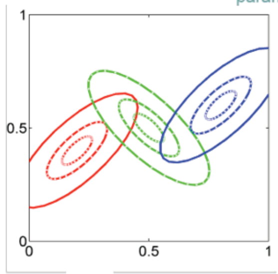
## Gaussian Mixture Models

- 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

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

*this is usually written as* $p(x|\Theta)$ *denoting the dependency on the parameters* $\Theta = \{\pi_k, \mu_k, \Sigma_k\}_{\{k \in \{1,2,\ldots,K\}\}}$

*Writing this as a conditional probability makes sense in connection with Bayesian Machine Learning (see [8])*

## 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

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

*this is usually written as* $p(x|\Theta)$ *denoting the dependency on the parameters* $\Theta = \{\pi_k, \mu_k, \Sigma_k\}_{\{k \in \{1,2,\ldots,K\}\}}$

*Writing this as a conditional probability makes sense in connection with Bayesian Machine Learning (see [8])*

# 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

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

*this is usually written as $p(x|\Theta)$ denoting the dependency on the parameters $\Theta = \{\pi_k, \mu_k, \Sigma_k\}_{\{k \in \{1,2,\dots,K\}\}}$*

*Writing this as a conditional probability makes sense in connection with Bayesian Machine Learning (see [8])*

0 0.5 1 [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

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

*this is usually written as $p(x|\Theta)$ denoting the dependency on the parameters $\Theta = \{\pi_k, \mu_k, \Sigma_k\}_{\{k \in \{1,2,\dots,K\}\}}$*

*Writing this as a conditional probability makes sense in connection with Bayesian Machine Learning (see [8])*

0 0.5 1 [6]


# Machine Learning

- For a distribution $p(x|\Theta)$ parametrized by a set of parameters $\Theta$ and iid data $X = \{x_1, x_2, \dots, x_N\}$, simple machine learning corresponds to finding the $\Theta$ 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 $\Theta$ 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 $\Theta$ and iid data $X = \{x_1, x_2, \dots, x_N\}$, simple machine learning corresponds to finding the $\Theta$ 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 $\Theta$ 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 $\Theta$ and iid data $X = \{x_1, x_2, \dots, x_N\}$ , simple machine learning corresponds to finding the $\Theta$ 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 $\Theta$ 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 $\Theta$ and iid data $X = \{x_1, x_2, \dots, x_N\}$ , simple machine learning corresponds to finding the $\Theta$ 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 $\Theta$ 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 $\Theta$ and iid data $X = \{x_1, x_2, \dots, x_N\}$ , simple machine learning corresponds to finding the $\Theta$ 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 $\Theta$ 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 $\Theta$ and iid data $X = \{x_1, x_2, \dots, x_N\}$ , simple machine learning corresponds to finding the $\Theta$ 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 $\Theta$ 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 $\Theta$ and iid data $X = \{x_1, x_2, \dots, x_N\}$ , simple machine learning corresponds to finding the $\Theta$ 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 $\Theta$ 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 $\Theta$ and iid data $X = \{x_1, x_2, \dots, x_N\}$ , simple machine learning corresponds to finding the $\Theta$ 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 $\Theta$ 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: $x \in \mathbb{R}^n$ 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:

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

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


## Machine Learning

Example: $x \in \mathbb{R}^n$ 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:

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

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

## Machine Learning

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

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

- 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:

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

$$\boldsymbol{\mu}_{\mathrm{ML}}: \frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}\,|\,\mu, \boldsymbol{\Sigma}) = 0 \qquad \boldsymbol{\Sigma}_{\mathrm{ML}}: \frac{\partial}{\partial \boldsymbol{\Sigma}} \ln p(\mathbf{X}\,|\,\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0$$

$$\Rightarrow \quad \boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \qquad \boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}$$


## Machine Learning

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

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

- 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:

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

$$\boldsymbol{\mu}_{\mathrm{ML}}: \frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}\,|\,\mu, \boldsymbol{\Sigma}) = 0 \qquad \boldsymbol{\Sigma}_{\mathrm{ML}}: \frac{\partial}{\partial \boldsymbol{\Sigma}} \ln p(\mathbf{X}\,|\,\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0$$

$$\Rightarrow \quad \boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \qquad \boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}$$



## Machine Learning

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

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

- 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:

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

$$\boldsymbol{\mu}_{\mathrm{ML}}: \frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}\,|\,\mu, \boldsymbol{\Sigma}) = 0 \qquad \boldsymbol{\Sigma}_{\mathrm{ML}}: \frac{\partial}{\partial \boldsymbol{\Sigma}} \ln p(\mathbf{X}\,|\,\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0$$

$$\Rightarrow \quad \boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \qquad \boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}$$



## Machine Learning

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

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

- 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:

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

$$\boldsymbol{\mu}_{\mathrm{ML}}: \frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}\,|\,\mu, \boldsymbol{\Sigma}) = 0 \qquad \boldsymbol{\Sigma}_{\mathrm{ML}}: \frac{\partial}{\partial \boldsymbol{\Sigma}} \ln p(\mathbf{X}\,|\,\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 0$$

$$\Rightarrow \quad \boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \qquad \boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}$$

## Machine Learning

Example: $x \in \mathbb{R}^n$ 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:

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

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


## Machine Learning

Example: $x \in \mathbb{R}^n$ 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:

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

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



## Machine Learning

Example: $x \in \mathbb{R}^n$ 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)

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

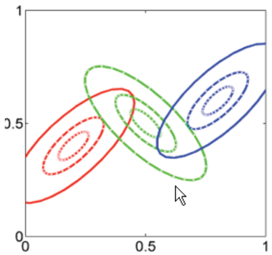
- Maximum log likelihood:

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

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

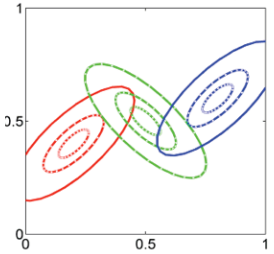

## GMM-Basics

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



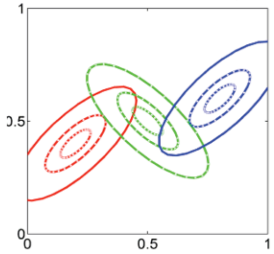
## GMM-Basics

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



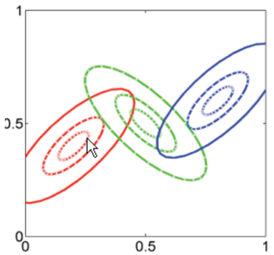

## GMM-Basics

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




## 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**

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

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

## GMM-Basics

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

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

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

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

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

- 1 of k representation

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

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

$p(z_k = 1) = \pi_k$

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

remark: If we have several observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$, then, because we have represented the marginal distribution in the form $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})$, it follows that for every observed data point $\mathbf{x}_n$ there is a corresponding latent variable $\mathbf{z}_n$.

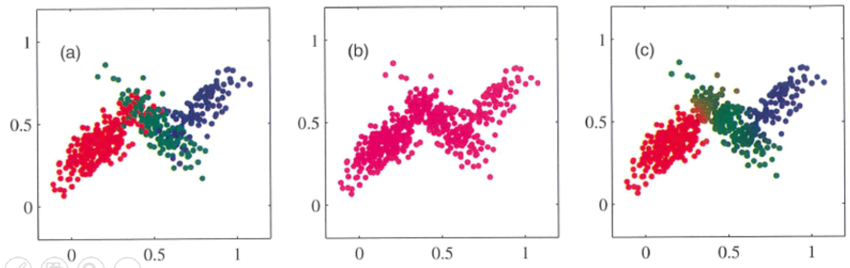
$$p(\mathbf{x}) = \sum_{\mathbf{z}} \underbrace{p(\mathbf{z})p(\mathbf{x}|\mathbf{z})}_{p(x,z)} = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



- 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

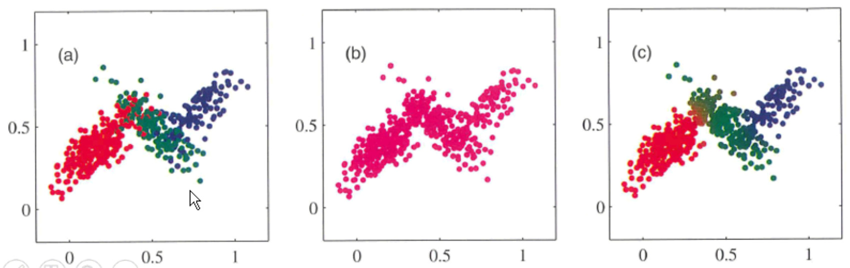



GMM-Basics

- 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

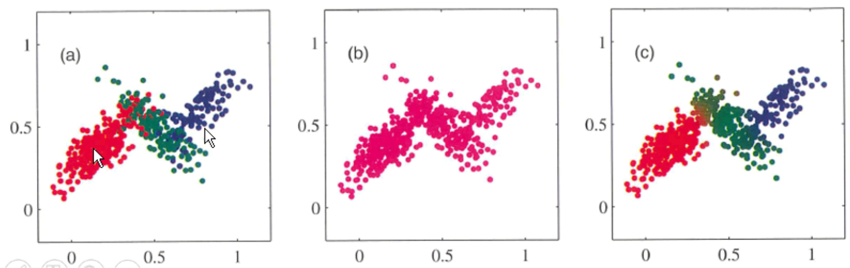



GMM-Basics

- 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

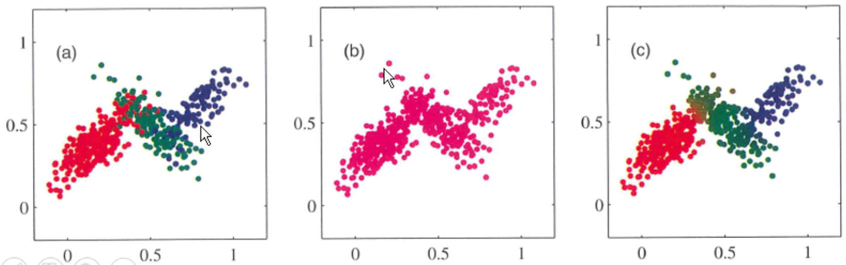


## GMM-Basics

- **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**



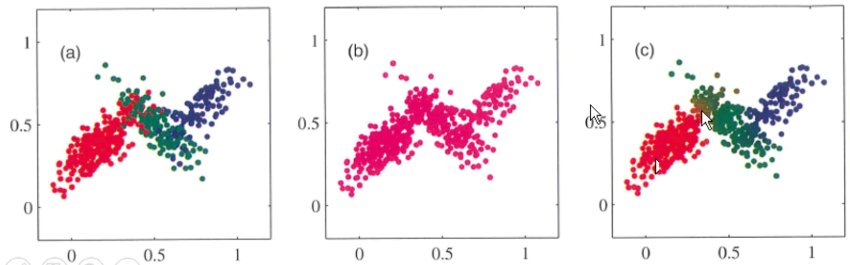


## GMM-Basics

- **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**

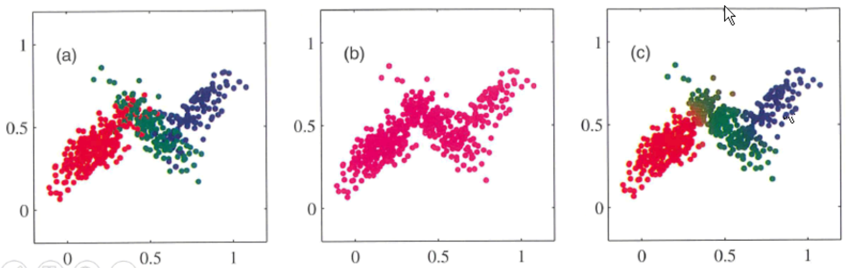


## GMM-Basics

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

- **1 of k representation**

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

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

$$p(z_k = 1) = \pi_k$$

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

remark: If we have several observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$, then, because we have represented the marginal distribution in the form $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})$, it follows that for every observed data point $\mathbf{x}_n$ there is a corresponding latent variable $\mathbf{z}_n$.

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

## GMM-Basics

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

- 1 of k representation

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

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

$p(z_k = 1) = \pi_k$

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

remark: If we have several observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$, then, because we have represented the marginal distribution in the form $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})$, it follows that for every observed data point $\mathbf{x}_n$ there is a corresponding latent variable $\mathbf{z}_n$.

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

## GMM-Basics

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 $\boldsymbol{\Sigma}$ →

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

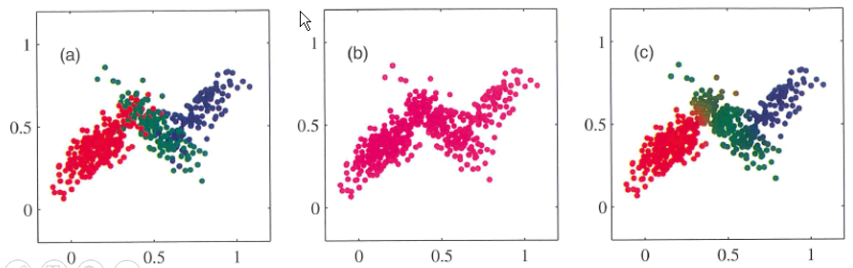
$$\left( N_k = \sum_{n=1}^{N} \gamma(z_{nk}). \right) \qquad \pi_k = \frac{N_k}{N}$$

## GMM-Basics

- 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 $\boldsymbol{\mu}_{\mathrm{k}}$

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

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

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

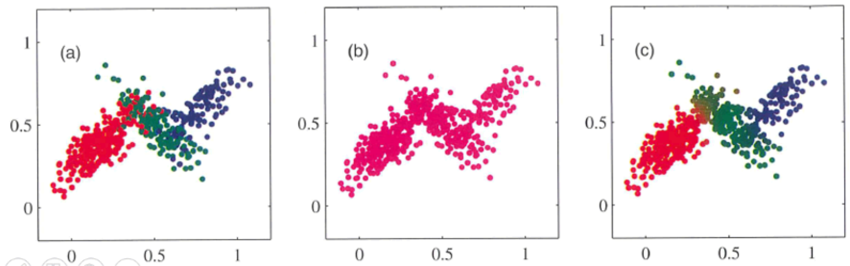
$$\left( N_k = \sum_{n=1}^{N}\gamma(z_{nk}). \right) \qquad \pi_k = \frac{N_k}{N}$$

## 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




## 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)**

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

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

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

- 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 $\gamma(z_{nk})_{(t-1)}$


## GMM-Basics

**Maximum likelihood (GMM)**

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

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

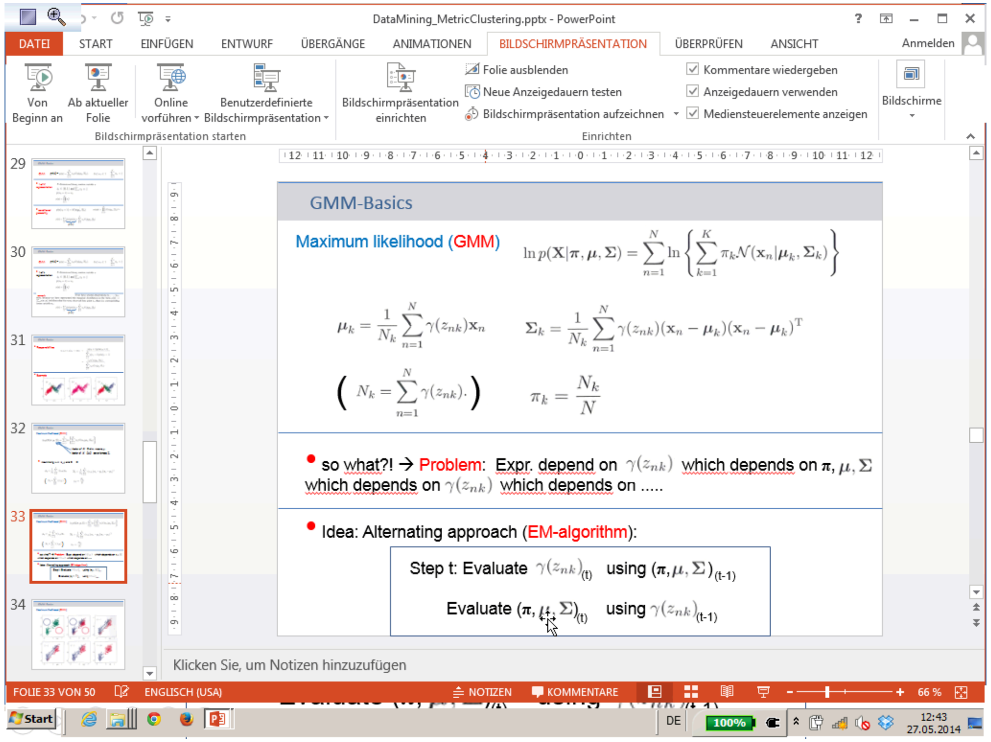
$$\left( N_k = \sum_{n=1}^{N}\gamma(z_{nk}). \right) \qquad \pi_k = \frac{N_k}{N}$$

- 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 $\gamma(z_{nk})_{(t-1)}$



## GMM-Basics

**Maximum likelihood (GMM)**

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

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

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

- 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 $\gamma(z_{nk})_{(t)}$

## GMM-Basics

Maximum likelihood (GMM)

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

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

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

- 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 $\gamma(z_{nk})_{(t)}$


## GMM-Basics

Maximum likelihood (GMM)

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

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

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

- 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 .....

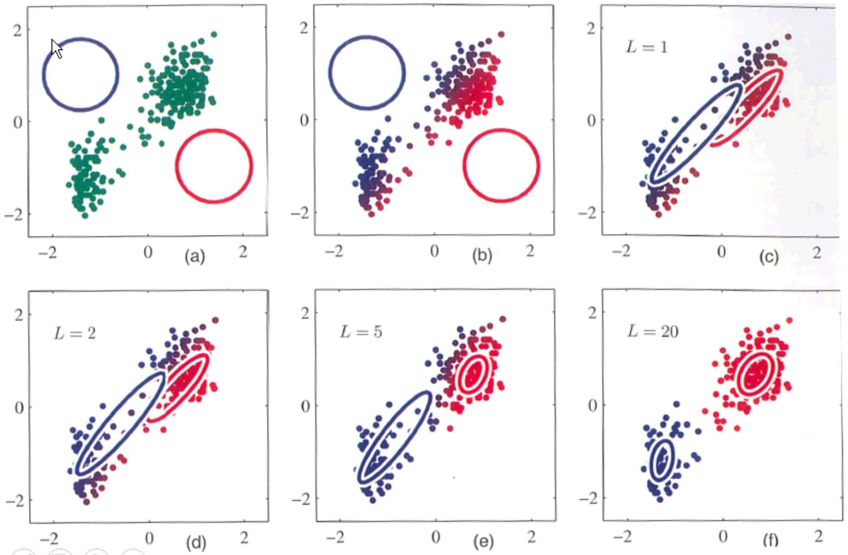
- Idea: Alternating approach (EM-algorithm):

  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 $\gamma(z_{nk})_{(t)}$


## GMM-Basics

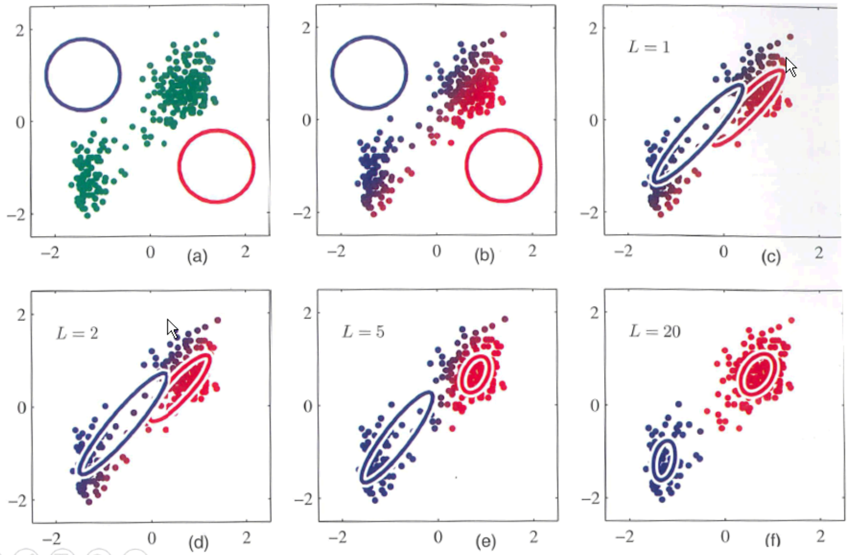
Maximum likelihood (GMM)





## GMM-Basics

Maximum likelihood (GMM)

## GMM-Basics

Maximum likelihood (GMM)

**EM for Gaussian Mixtures**

Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters (comprising the means and covariances of the components and the mixing coefficients).

1. Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients $\pi_k$, and evaluate the initial value of the log likelihood.
2. **E step**. Evaluate the responsibilities using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{9.23}$$

## 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 $\theta$ 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

- 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 $\theta$ 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

- 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 $\theta$ 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

- 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 $\theta$ 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

- 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 $\theta$ 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 $\ln p(\mathbf{X},\mathbf{Z}|\theta)$


## GMM-Basics

### Maximum likelihood (GMM)

#### EM for Gaussian Mixtures

Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters (comprising the means and covariances of the components and the mixing coefficients).

1. Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients $\pi_k$, and evaluate the initial value of the log likelihood.
2. **E step**. Evaluate the responsibilities using the current parameter values

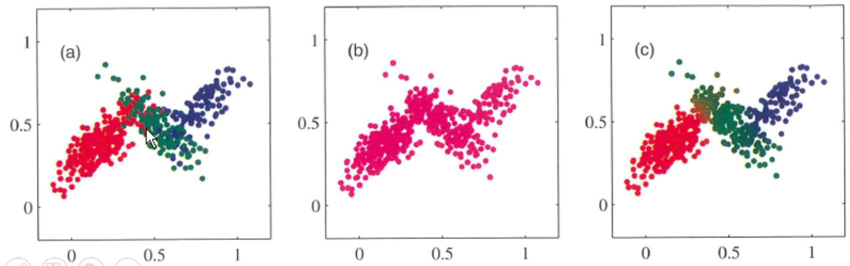
$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}. \tag{9.23}$$

## 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



## 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 $\theta$ 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

- 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 $\theta$ 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: Relation to K-Means

*that is why K-Means favors spherical clusters*

- If we use k Gaussians with Σ = εI:

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

- 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

*that is why K-Means favors spherical clusters*

- If we use k Gaussians with Σ = εI:

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

- 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

*that is why K-Means favors spherical clusters*

- If we use k Gaussians with Σ = εI:

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

- 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

*that is why K-Means favors spherical clusters*

- If we use k Gaussians with Σ = εI:

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

- 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


## Real World Networks: Properties and Models

Lecture will mostly follow [1], thus corresponding citations are often omitted to increase readability

## 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

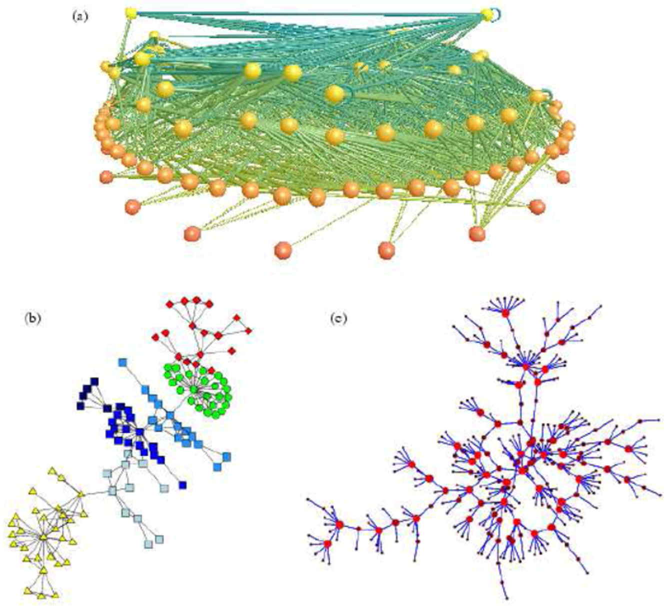



FIG. 2 Three examples of the kinds of networks that are the topic of this review. (a) A food web of predator-prey interactions between species in a freshwater lake [272]. Picture courtesy of Neo Martinez and Richard Williams. (b) The network of collaborations between scientists at a private research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

[1]

## Mean Average Path Length

- "Small World Effect": $l(n)$ "small" → $l(n) \in O(log(n))$
- undirected graph:

$$\ell = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}$$

formula also counts 0 distances from i to i: ½ n(n+1) = ½ n(n-1) + n

- Expression allowing for disconnected components (where $d_{ij}=\infty$ can occur): harmonic mean:

$$\ell^{-1} = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}^{-1}$$

## Mean Average Path Length

- "Small World Effect": $l(n)$ "small" → $l(n) \in O(log(n))$
- undirected graph:

$$\ell = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}$$

formula also counts 0 distances from i to i: ½ n(n+1) = ½ n(n-1) + n

- Expression allowing for disconnected components (where $d_{ij}=\infty$ can occur): harmonic mean:

$$\ell^{-1} = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}^{-1}$$

## Mean Average Path Length

- "Small World Effect": *l(n)* "small" → $l(n) \in O(\log(n))$
- undirected graph:

$$\ell = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}$$

formula also counts 0 distances from i to i: ½ n(n+1) = ½ n(n-1) + n

- Expression allowing for disconnected components (where $d_{ij}=\infty$ can occur): harmonic mean:

$$\ell^{-1} = \frac{1}{\frac{1}{2}n(n+1)} \sum_{i \geq j} d_{ij}^{-1}$$

## Transitivity / Clustering Coefficient

- Clustering coefficient (whole graph):

$$C = C^{(1)} = \frac{3\times \text{ number of triangles in the network}}{\text{number of connected triples of vertices}} = \frac{6\times \text{ number of triangles in the network}}{\text{number of paths of length two}}$$

p(FOAF)

- Clustering coefficient (Watts-Strogatz-version, for node i):

$$C_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered on vertex } i} = \frac{|\{e_{\{kj\}} \mid v_k, v_j \in N_i\}|}{\frac{k_i(k_i-1)}{2}}$$

(see Introduction , $k_i$ = degree of node i)

Clustering coefficient (Watts-Strogatz-version, for whole graph):

$$C = C^{(2)} = \frac{1}{n}\sum_i C_i$$

mean of ratio instead of ratio of means

## Transitivity / Clustering Coefficient

- Clustering coefficient (whole graph):

$$C = C^{(1)} = \frac{3\times \text{ number of triangles in the network}}{\text{number of connected triples of vertices}} = \frac{6\times \text{ number of triangles in the network}}{\text{number of paths of length two}}$$

p(FOAF)

- Clustering coefficient (Watts-Strogatz-version, for node i):

$$C_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered on vertex } i} = \frac{|\{e_{\{kj\}} \mid v_k, v_j \in N_i\}|}{\frac{k_i(k_i-1)}{2}}$$

(see Introduction , $k_i$ = degree of node i)

Clustering coefficient (Watts-Strogatz-version, for whole graph):

$$C = C^{(2)} = \frac{1}{n}\sum_i C_i$$

mean of ratio instead of ratio of means

## Transitivity / Clustering Coefficient

- Clustering coefficient (whole graph):

$$C = C^{(1)} = \frac{3\times \text{ number of triangles in the network}}{\text{number of connected triples of vertices}} = \frac{6\times \text{ number of triangles in the network}}{\text{number of paths of length two}}$$

p(FOAF)

- Clustering coefficient (Watts-Strogatz-version, for node i):

$$C_i = \frac{\text{number of triangles connected to vertex } i}{\text{number of triples centered on vertex } i} = \frac{|\{e_{\{kj\}} \mid v_k, v_j \in N_i\}|}{\frac{k_i(k_i-1)}{2}}$$

(see Introduction , $k_i$ = degree of node i)

Clustering coefficient (Watts-Strogatz-version, for whole graph):

$$C = C^{(2)} = \frac{1}{n}\sum_i C_i$$

mean of ratio instead of ratio of means