CS559 Machine Learning
EM and Latent Variable Models

Tian Han

Department of Computer Science
Stevens Institute of Technology

Week 10

Outline

- Discrete Latent Variable Model
- An Alternative View of EM
- Continuous Latent Variable Model

Latent Variable Model

Latent variable models

The model considers *unobserved/missing/hidden* values can be important, especially in **unsupervised learning**.

Latent variable models

The model considers *unobserved/missing/hidden* values can be important, especially in **unsupervised learning**.

- Offer a lower dimensional hidden representation of the data and their dependencies.

Latent variable models

The model considers *unobserved/missing/hidden* values can be important, especially in **unsupervised learning**.

- Offer a lower dimensional hidden representation of the data and their dependencies.
- Real dataset may have missing/corrupted values.

Latent variable models

The model considers *unobserved/missing/hidden* values can be important, especially in **unsupervised learning**.

- Offer a lower dimensional hidden representation of the data and their dependencies.
- Real dataset may have missing/corrupted values.

→ **Latent Variable Model**, $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$

Figure: [C. Bishop, PRML]

GMM as discrete latent variable model

Recall *Gaussian Mixture Model* can be written as:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

- We introduce a K-dimensional binary random variable $\mathbf{z}$ having a 1-of-K representation in which a particular element $z_k = 1$ and all other elements are equal to 0.
  Thus, $z_k \in \{0, 1\}, \sum_k z_k = 1$
- The marginal distribution over $\mathbf{z}$:

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

- Write the distribution in this form:

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

GMM as discrete latent variable model

- The conditional distribution of $\mathbf{x}$ given a particular value for $\mathbf{z}$ is a Gaussian:

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)^{z_k}$$

- The marginal distribution of $\mathbf{x}$ is then obtained by summing the joint distribution over all possible values of $\mathbf{z}$:

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$

Posterior/responsibility

- The posterior probabilities $p(k|\mathbf{x})$ (responsibilities):

$$\gamma_k(\mathbf{x}) = p(z_k = 1|\mathbf{x}) = \gamma(z_{nk}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_j p(z_j = 1)p(\mathbf{x}|z_j = 1)}$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j)}$$

- The parameters: π, μ, Σ

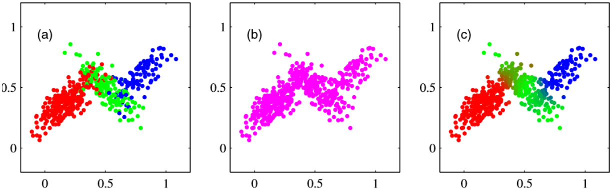

Figure: (a): *complete* data $(\mathbf{x}, \mathbf{z})$. (b): *incomplete* data $\mathbf{x}$. (c): posterior/responsibilities $\gamma(z_k)$. [C. Bishop, PRML]

Graphical representation

Graphical representation of a GMM for a set of N i.i.d data points $\{\mathbf{x}_n\}$, with corresponding latent variables $\{\mathbf{z}_n\}$, where $n = 1, 2, \ldots, N$

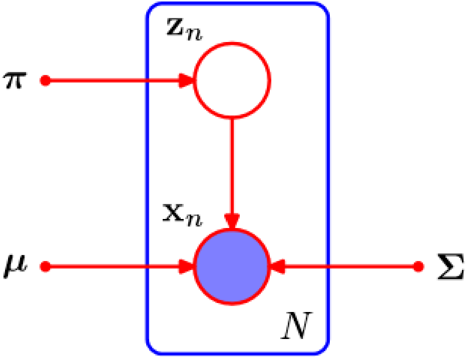

Figure: [C. Bishop, PRML]

Finding maximum likelihood solution using EM

Now we re-write the MLE solutions in last lecture using discrete latent variable $\mathbf{z}$.

For μ_k

- Setting the derivatives of $\ln p(\mathbf{X}|\pi, \mu, \Sigma)$ w.r.t μ_k to 0:

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j)}}_{\gamma(z_{nk})} \Sigma_k(\mathbf{x}_n - \mu_k)$$

- We get:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$

For Σ_k

- Setting the derivatives of $\ln p(\mathbf{X}|\pi, \mu, \Sigma)$ w.r.t Σ_k to 0
- We get:

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$

For π_k

- Finally, we maximize $\ln p(\mathbf{X}|\pi, \mu, \Sigma)$ w.t.t π_k
- We know that $\sum_k \pi_k = 1$
- Using a Lagrange multiplier and maximizing the following quantity:

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) + \lambda(\sum_{k=1}^{K} \pi_k - 1)$$

- We get:

$$\pi_k = \frac{N_k}{N}$$

EM algorithm: initialization

- Initialize the means μ_k, covariances Σ_k and mixing coefficient π_k, and evaluate the initial value of the log likelihood.

EM algorithm: E step

- **E step**. Evaluate the responsibilities using the current parameter values:

$$\gamma_k(\mathbf{x}) = \gamma(z_{nk}) = p(z_k = 1|\mathbf{x}) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j)}$$

EM algorithm: M step

- **M step**. Re-estimate the parameters using the current responsibilities (where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$)

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

$$\pi_k = \frac{N_k}{N}$$

EM algorithm: Check convergence

- Evaluate the log likelihood

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \Big\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k) \Big\}$$

check for convergence of either the parameters or the log likelihood.

EM for Latent Variable Model

Notations

- Denote the set of all observed data by $\mathbf{X}$, in which the n-th row represents $\mathbf{x}_n^T$, and all data are assumed to be i.i.d.
- Similarly denote the set of all latent variables by $\mathbf{Z}$, with a corresponding row $\mathbf{z}_n^T$.
- The set of all model parameters are denoted by θ.

Maximum likelihood is not easy

For latent variable model, the likelihood of $\mathbf{X}$ (the marginal distribution of $\mathbf{X}$) is:

$$p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$$

Maximum likelihood is not easy

For latent variable model, the likelihood of $\mathbf{X}$ (the marginal distribution of $\mathbf{X}$) is:

$$p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$$

Therefore, the log likelihood function is given by:

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

Maximum likelihood is not easy

For latent variable model, the likelihood of $\mathbf{X}$ (the marginal distribution of $\mathbf{X}$) is:

$$p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$$

Therefore, the log likelihood function is given by:

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

- The discussion will apply equally well to *continuous latent variables* $\rightarrow$ replace the sum over $\mathbf{Z}$ with an integral.

Maximum likelihood is not easy

For latent variable model, the likelihood of $\mathbf{X}$ (the marginal distribution of $\mathbf{X}$) is:

$$p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$$

Therefore, the log likelihood function is given by:

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

- The discussion will apply equally well to *continuous latent variables* $\rightarrow$ replace the sum over $\mathbf{Z}$ with an integral.
- The summation over the latent variables appears inside the logarithm $\rightarrow$ resulting complicated ML solution. (even if joint $p(\mathbf{X}, \mathbf{Z}, |\theta)$ belongs to exponential family, the marginal $p(\mathbf{X}|\theta)$ typically does not.)

Complete data vs incomplete data

- *Complete data set*: $\{\mathbf{X}, \mathbf{Z}\}$, each observation and corresponding latent variable.
- *Incomplete data set*: $\mathbf{X}$, actual observed data.

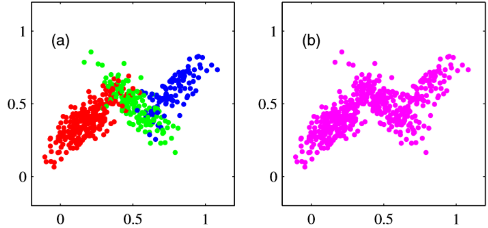

Figure: (a): *complete* data $(\mathbf{x}, \mathbf{z})$. (b): *incomplete* data $\mathbf{x}$. [C. Bishop, PRML]

- **Assume**: maximization of complete-data log likelihood function, i.e., $\ln p(\mathbf{X}, \mathbf{Z}|\theta)$, is straightforward.

Infer the latent variables

- In practice, we are not given the complete data set $\{\mathbf{X}, \mathbf{Z}\}$, but only the incomplete data $\mathbf{X}$.

Infer the latent variables

- In practice, we are not given the complete data set $\{\mathbf{X}, \mathbf{Z}\}$, but only the incomplete data $\mathbf{X}$.
- Our state of knowledge of the values of the latent variables $\mathbf{Z}$ is given only by the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \theta)$.

Infer the latent variables

- In practice, we are not given the complete data set $\{\mathbf{X}, \mathbf{Z}\}$, but only the incomplete data $\mathbf{X}$.
- Our state of knowledge of the values of the latent variables $\mathbf{Z}$ is given only by the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \theta)$.
- Cannot use the complete-data log likelihood, consider instead its expected value under the posterior distribution of the latent variable. (i.e., consider $\mathbb{E}_{\mathbf{Z}^* \sim p(\mathbf{Z}|\mathbf{X},\theta)}[\ln p(\mathbf{X}, \mathbf{Z}^*|\theta)]$

Infer latent variables and update θ

- Suppose the current parameter value θ^{old}.

Infer latent variables and update θ

- Suppose the current parameter value θ^{old}.
- Find posterior distribution of latent variables $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$.

Infer latent variables and update θ

- Suppose the current parameter value θ^{old}.
- Find posterior distribution of latent variables $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$.
- Use this posterior distribution to find the expectation of the complete-data log likelihood evaluated for some general parameter value θ.

Infer latent variables and update θ

- Suppose the current parameter value θ^{old}.
- Find posterior distribution of latent variables $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$.
- Use this posterior distribution to find the expectation of the complete-data log likelihood evaluated for some general parameter value θ.
- Such expectation is denoted as $\mathcal{Q}(\theta, \theta^{old})$:

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

  - Logarithm acts directly on the joint distribution $p(\mathbf{X}, \mathbf{Z}|\theta)$

Infer latent variables and update θ

- Suppose the current parameter value θ^{old}.
- Find posterior distribution of latent variables $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$.
- Use this posterior distribution to find the expectation of the complete-data log likelihood evaluated for some general parameter value θ.
- Such expectation is denoted as $\mathcal{Q}(\theta, \theta^{old})$:

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

  - Logarithm acts directly on the joint distribution $p(\mathbf{X}, \mathbf{Z}|\theta)$

- We then determine the revised parameter estimate θ^{new} by maximizing this function:

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

The general EM algorithm

Goal: maximize the likelihood function $p(\mathbf{X}|\theta)$ with respect to θ.
Given joint distribution $p(\mathbf{X}, \mathbf{Z}|\theta)$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters θ.

The general EM algorithm

Goal: maximize the likelihood function $p(\mathbf{X}|\theta)$ with respect to θ.
Given joint distribution $p(\mathbf{X}, \mathbf{Z}|\theta)$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters θ.

1. Choose an initial setting for parameter θ^{old}.

The general EM algorithm

Goal: maximize the likelihood function $p(\mathbf{X}|\theta)$ with respect to θ.
Given joint distribution $p(\mathbf{X}, \mathbf{Z}|\theta)$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters θ.

1. Choose an initial setting for parameter θ^{old}.
2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$.

The general EM algorithm

Goal: maximize the likelihood function $p(\mathbf{X}|\theta)$ with respect to θ.
Given joint distribution $p(\mathbf{X}, \mathbf{Z}|\theta)$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters θ.

1. Choose an initial setting for parameter θ^{old}.
2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$.
3. **M step** Evaluate θ^{new} given by:

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

where

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

The general EM algorithm

Goal: maximize the likelihood function $p(\mathbf{X}|\theta)$ with respect to θ.
Given joint distribution $p(\mathbf{X}, \mathbf{Z}|\theta)$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters θ.

1. Choose an initial setting for parameter θ^{old}.
2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$.
3. **M step** Evaluate θ^{new} given by:

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

   where

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

4. Check for convergence of either log likelihood or parameter values. If not converged, let:

$$\theta^{old} \leftarrow \theta^{new}$$

   and return to step 2.

Gaussian Mixture revisited

- Gaussian Mixture Model can be seen as a latent variable model where latent variable is discrete. Can we use the general EM algorithm to get MLE?

Gaussian Mixture revisited

- Gaussian Mixture Model can be seen as a latent variable model where latent variable is discrete. Can we use the general EM algorithm to get MLE?
- The general EM deals with the joint distribution $p(\mathbf{X}, \mathbf{Z}|\theta)$. So suppose that in addition to the observed data set $\mathbf{X}$, we were also given the values of the corresponding discrete variables $\mathbf{Z}$.

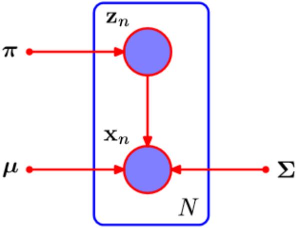

Figure: Suppose discrete variables $\mathbf{z}_n$ are observed as well as data $\mathbf{x}_n$ [C. Bishop, PRML]

Maximize the likelihood for complete data $\{\mathbf{X}, \mathbf{Z}\}$

Recall the prior on discrete latent variable:

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

The conditional:

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)^{z_k}$$

The likelihood for joint using i.i.d assumption:

$$p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)^{z_{nk}}$$

Maximize the likelihood for complete data $\{\mathbf{X}, \mathbf{Z}\}$

Taking logarithm of previous likelihood function:

$$\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)\}$$

Maximize the likelihood for complete data $\{\mathbf{X}, \mathbf{Z}\}$

Taking logarithm of previous likelihood function:

$$\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)\}$$

Recall the log-likelihood function we derive directly for Gaussian Mixture Model:

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \Big\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k) \Big\}$$

Maximize the likelihood for complete data $\{\mathbf{X}, \mathbf{Z}\}$

Taking logarithm of previous likelihood function:

$$\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)\}$$

Recall the log-likelihood function we derive directly for Gaussian Mixture Model:

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \Big\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k) \Big\}$$

- For log-likelihood on incomplete data $\mathbf{X}$, summation over K is inside the logarithm.

Maximize the likelihood for complete data $\{\mathbf{X}, \mathbf{Z}\}$

Taking logarithm of previous likelihood function:

$$\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)\}$$

Recall the log-likelihood function we derive directly for Gaussian Mixture Model:

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

- For log-likelihood on incomplete data $\mathbf{X}$, summation over K is inside the logarithm.
- For log-likelihood on complete data $\{\mathbf{X}, \mathbf{Z}\}$, logarithm act directly on Gaussian distribution $\rightarrow$ easy to maximize w.r.t parameters.

Infer latent variable from the posterior distribution

- Complete data log-likelihood can be maximized easily and trivially in closed form.

Infer latent variable from the posterior distribution

- Complete data log-likelihood can be maximized easily and trivially in closed form.
- However, in practice, we **do not** have values for the latent variables.

Infer latent variable from the posterior distribution

- Complete data log-likelihood can be maximized easily and trivially in closed form.
- However, in practice, we **do not** have values for the latent variables.
- Consider the expectation, with respect to the posterior distribution of the latent variables, of the complete-data log-likelihood.

Posterior of the latent variable

Recall the prior on discrete latent variable:

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

The conditional:

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)^{z_k}$$

Use Bayes' theorem, the posterior of the latent is given by:

$$p(\mathbf{Z}|\mathbf{X}, \mu, \Sigma, \pi) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)]^{z_{nk}}$$

Expectation of complete-data log-likelihood

Recall the form of the complete-data log-likelihood is:

$$\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)\}$$

Expectation of complete-data log-likelihood

Recall the form of the complete-data log-likelihood is:

$$\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)\}$$

Since latent variables z_{nk} are unknown, we need to consider the expectation of $\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi)$ under the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \mu, \Sigma, \pi)$, i.e., $\mathbb{E}_{\mathbf{Z} \sim p(\mathbf{Z}|\mathbf{X}, \mu, \Sigma, \pi)}[\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi)]$.

Expectation of complete-data log-likelihood

Recall the form of the complete-data log-likelihood is:

$$\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk}\{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)\}$$

Since latent variables z_{nk} are unknown, we need to consider the expectation of $\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi)$ under the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \mu, \Sigma, \pi)$, i.e., $\mathbb{E}_{\mathbf{Z}\sim p(\mathbf{Z}|\mathbf{X},\mu,\Sigma,\pi)}[\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi)]$.

We only need to get expected value of the indicator variable z_{nk} under posterior $p(\mathbf{Z}|\mathbf{X}, \mu, \Sigma, \pi)$, i.e, (you could verify this!)

$$\mathbb{E}[z_{nk}] = \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j)} = \gamma(z_{nk})$$

- This is just the responsibility of component k for data point $\mathbf{x}_n$.

Expectation of complete-data log-likelihood

The expected value of the complete-data log likelihood function is given by:

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi)] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)\}$$

General EM for GMM

1. Choose an initial setting for parameter $\mu^{old}, \Sigma^{old}, \pi^{old}$.

General EM for GMM

1. Choose an initial setting for parameter $\mu^{old}, \Sigma^{old}, \pi^{old}$.
2. **E step** Evaluate $\mathbb{E}[z_{nk}] = \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j)} = \gamma(z_{nk})$

General EM for GMM

1. Choose an initial setting for parameter $\mu^{old}, \Sigma^{old}, \pi^{old}$.
2. **E step** Evaluate $\mathbb{E}[z_{nk}] = \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j)} = \gamma(z_{nk})$
3. **M step** Evaluate $\mu^{new}, \Sigma^{new}, \pi^{new}$ by maximizing $\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi)]$, i.e.,

$$
\begin{aligned}
\mu_k &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \\
\Sigma_k &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \\
\pi_k &= \frac{N_k}{N}
\end{aligned}
$$

General EM for GMM

1. Choose an initial setting for parameter $\mu^{old}, \Sigma^{old}, \pi^{old}$.
2. **E step** Evaluate $\mathbb{E}[z_{nk}] = \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\mu_j, \Sigma_j)} = \gamma(z_{nk})$
3. **M step** Evaluate $\mu^{new}, \Sigma^{new}, \pi^{new}$ by maximizing $\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi)]$, i.e.,

$$
\begin{aligned}
\mu_k &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \\
\Sigma_k &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \\
\pi_k &= \frac{N_k}{N}
\end{aligned}
$$

4. Check for convergence. If not, return to step 2 by letting:

$$\{\mu, \Sigma, \pi\}^{old} \leftarrow \{\mu, \Sigma, \pi\}^{new}$$

(Same as EM algorithm for GMM we derived in last lecture!)

Continuous Latent Variable Model

Continuous Latent Variable

- In mixture models, we have discrete latent variable z:
  - E.g., in mixture of Guassians, $p(z_k = 1) = \pi_k$, and $p(x|z)$ is Gaussian

Continuous Latent Variable

- In mixture models, we have discrete latent variable z:
  - E.g., in mixture of Guassians, $p(z_k = 1) = \pi_k$, and $p(x|z)$ is Gaussian
- We could also have continuous latent variable z!

Continuous Latent Variable

- In mixture models, we have discrete latent variable z:
  - E.g., in mixture of Guassians, $p(z_k = 1) = \pi_k$, and $p(x|z)$ is Gaussian
- We could also have continuous latent variable z!
- Why do we need a continuous z?

Low dimensional manifold

Many data points lie close to a manifold of much lower dimensionality than that of the original data space.

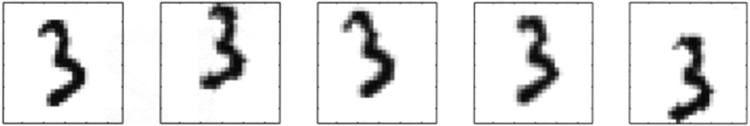

Figure: There are only three degrees of freedom of variability: vertical and horizontal translation and rotations.[C. Bishop, PRML]

Generative view

- In practice, the data points will not be confined precisely to a smooth low dimensional manifold, and we can interpret the departures of data points from the manifold as “noise”.

Generative view

- In practice, the data points will not be confined precisely to a smooth low dimensional manifold, and we can interpret the departures of data points from the manifold as “noise”.
- Naturally lead to **generative view** of the model:
  - First select a point within the manifold according to some latent variable distribution $p(z)$
  - Then generate an observed data point by adding noise, drawn from some conditional distribution of the data variables given the latent variables $p(x|z)$.

Generative view

- In practice, the data points will not be confined precisely to a smooth low dimensional manifold, and we can interpret the departures of data points from the manifold as "noise".
- Naturally lead to **generative view** of the model:
  - First select a point within the manifold according to some latent variable distribution $p(z)$
  - Then generate an observed data point by adding noise, drawn from some conditional distribution of the data variables given the latent variables $p(x|z)$.
- *Linear Gaussian* latent variable model $\rightarrow$ PCA, factor analysis etc

Probabilistic PCA

- Probabilistic, generative view of data.[Tipping and Bishop (1997), Roweis (1998)]
- Simple example of linear-Gaussian framework, in which all of the marginal and conditional distributions are Gaussian.
- Introduce an explicit latent variable $\mathbf{z}$ corresponding to the *principal-component subspace*, define the Gaussian prior and conditional:

$$
\begin{aligned}
p(\mathbf{z}) &= \mathcal{N}(\mathbf{z}|0, \mathbf{I}) \\
p(\mathbf{x}|\mathbf{z}) &= \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \mu, \sigma^2 \mathbf{I})
\end{aligned}
$$

Probabilistic PCA

- Probabilistic, generative view of data.[Tipping and Bishop (1997), Roweis (1998)]
- Simple example of linear-Gaussian framework, in which all of the marginal and conditional distributions are Gaussian.
- Introduce an explicit latent variable $\mathbf{z}$ corresponding to the *principal-component subspace*, define the Gaussian prior and conditional:

$$\begin{aligned} p(\mathbf{z}) &= \mathcal{N}(\mathbf{z}|0, \mathbf{I}) \\ p(\mathbf{x}|\mathbf{z}) &= \mathcal{N}(\mathbf{x}|\mathbf{W}\mathbf{z} + \mu, \sigma^2\mathbf{I}) \end{aligned}$$

- Equivalently: (note that $\mathbf{x} \in \mathcal{R}^D$, and $\mathbf{z} \in \mathcal{R}^M$)

$$\begin{aligned} \mathbf{x} &= \mathbf{W}\mathbf{z} + \mu + \epsilon \\ \mathbf{z} &\sim \mathcal{N}(0, \mathbf{I}), \epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}) \end{aligned}$$

Illustration

Generative view: an observed data point $\mathbf{x}$ is generated by:

- draw a value $\hat{z}$ for the latent variable from $p(\mathbf{z})$.
- draw a value for $\mathbf{x}$ from isotropic Gaussian having mean $\mathbf{Wz} + \mu$ and covariance $\sigma^2\mathbf{I}$.

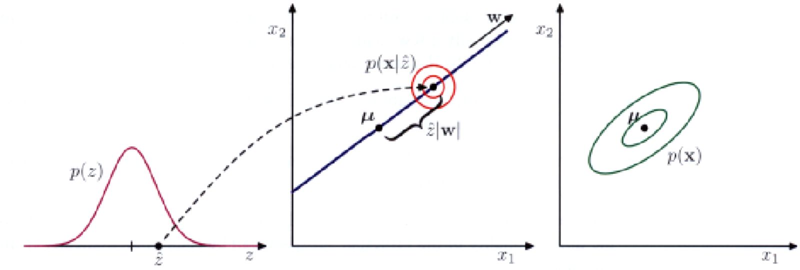

Figure: Illustrates the generative view of Probabilistic PCA for 2D data. [C. Bishop, PRML]

Marginal distribution

- The marginal distribution $p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$ is **Gaussian**:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \mathcal{C})$$

where covariance matrix $\mathcal{C}$:

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

Posterior distribution

- The posterior distribution $p(\mathbf{z}|\mathbf{x})$ is again **Gaussian**:

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

where

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

- Posterior mean depends on $\mathbf{x}$, the posterior covariance is independent of $\mathbf{x}$.

MLE for Probabilistic PCA

Maximize the following log-likelihood w.r.t parameters $\mu, \mathbf{W}, \sigma^2$:

$$\ln p(\mathbf{X}|\mu, \mathbf{W}, \sigma^2) = \sum_{n=1}^{N} \ln p(\mathbf{x}_n|\mu, \mathbf{W}, \sigma^2)$$

- Closed form (but complicated) solutions.
- Could also use EM algorithm. (Offer **computational efficiency** in spaces of high-dimensionality)

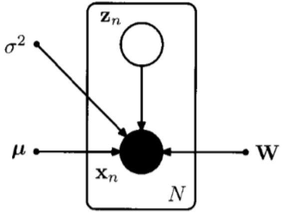

Figure: Probabilistic PCA [C. Bishop, PRML]

Probabilistic PCA and conventional PCA

- Conventional PCA is generally formulated as a projection of points from the D-dimensional data space onto an M-dimensional linear subspace

Probabilistic PCA and conventional PCA

- Conventional PCA is generally formulated as a projection of points from the D-dimensional data space onto an M-dimensional linear subspace
- Probabilistic PCA is most naturally expressed as a generative mapping from the M-dimensional latent space into the D-dimensional data space.

Probabilistic PCA and conventional PCA

- Conventional PCA is generally formulated as a projection of points from the D-dimensional data space onto an M-dimensional linear subspace
- Probabilistic PCA is most naturally expressed as a generative mapping from the M-dimensional latent space into the D-dimensional data space.
- As $\sigma^2 \to 0$, it can be shown to recover the conventional PCA. (see PRML 12.2)

Probabilistic PCA and conventional PCA

- Conventional PCA is generally formulated as a projection of points from the D-dimensional data space onto an M-dimensional linear subspace
- Probabilistic PCA is most naturally expressed as a generative mapping from the M-dimensional latent space into the D-dimensional data space.
- As $\sigma^2 \to 0$, it can be shown to recover the conventional PCA. (see PRML 12.2)
- Having a fully probabilistic model for PCA, we could deal with missing data, and can be naturally treated using EM algorithm.

Probabilistic PCA and conventional PCA

- Conventional PCA is generally formulated as a projection of points from the D-dimensional data space onto an M-dimensional linear subspace
- Probabilistic PCA is most naturally expressed as a generative mapping from the M-dimensional latent space into the D-dimensional data space.
- As $\sigma^2 \to 0$, it can be shown to recover the conventional PCA. (see PRML 12.2)
- Having a fully probabilistic model for PCA, we could deal with missing data, and can be naturally treated using EM algorithm.
- We can also consider different distribution for $\mathbf{x}|\mathbf{z}$:
  - E.g., Laplace distribution if you want it to be robust.
  - E.g., logistic or softmax if you have discrete $\mathbf{x}$

Acknowledgement and Further Reading

Part of slides are taken from Dr. Y. Ning's Spring 19 offering of CS-559.

Part of slides are inspired by CPSC540: Machine Learning by Mark Schmidt.

Further Reading:
Chapter 9.2, 9.3, 12.2 of *Pattern Recognition and Machine Learning* by C. Bishop.