# Likelihood and Sampling

Morteza Haghir Chehreghani

March 31, 2011

# Overview

# Maximum Likelihood Estimation (MLE)

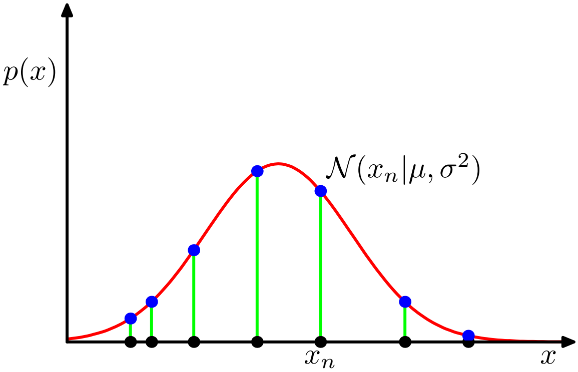



Figure: Illustration of the likelihood function for a Gaussian distribution, shown by the red curve. Here the black points denote a data set of values $\{x_n\}$, and the likelihood function $p(\mathbf{X}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$ corresponds to the product of the blue values. Maximizing the likelihood involves adjusting the mean and variance of the Gaussian so as to maximize this product.

# MLE for Gaussian – Univariate Case 1/3

Data likelihood:

$$p(\mathbf{X}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2).$$

Taking the logarithm and inserting the Gaussian distribution:

$$\ln p(\mathbf{X}|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi).$$

Which we want to maximize w.r.t. $\mu$ and $\sigma$ in order to maximize the probability of the observed data, given the Gaussian model.

# MLE for Gaussian – Univariate Case 2/3

Taking the derivative w.r.t. $\mu$:

$$\frac{\partial \ln p(\mathbf{X}|\mu, \sigma)}{\partial \mu} = \frac{1}{2\sigma^2} \sum_{n=1}^{N} 2(x_n - \mu).$$

Setting this to zero leads to the MLE for the mean parameter:

$$\widehat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n.$$

# MLE for Gaussian – Univariate Case 3/3

Taking the derivative w.r.t. $\sigma^2$:

$$\frac{\partial \ln p(\mathbf{X}|\mu, \sigma)}{\partial \sigma^2} = \frac{1}{2}\sigma^{-4} \sum_{n=1}^{N} (x_n - \mu) - \frac{N}{2}\sigma^{-2}.$$

Setting this to zero leads to the MLE for the mean parameter:

$$\widehat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu)^2.$$

Inserting $\widehat{\mu}$ for $\mu$:

$$\widehat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \widehat{\mu})^2.$$

# MLE for Gaussian – Multivariate Case

Slightly trickier as we have to take derivatives w.r.t. vectors. Here only for $\boldsymbol{\mu}$. The log likelihood is given by

$$\ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln\left(\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\right) - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}).$$

Expanding the square and taking the derivative w.r.t. $\boldsymbol{\mu}$ we get

$$\frac{\partial}{\partial \boldsymbol{\mu}} \sum_{n=1}^{N} \left(\mathbf{x}_n^T \boldsymbol{\Sigma}^{-1}\mathbf{x}_n - 2\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{x}_n + \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right) \stackrel{!}{=} \mathbf{0}$$

which leads to

$$\sum_{n=1}^{N} -2\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_n + 2N\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \stackrel{!}{=} \mathbf{0},$$

and thus

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n.$$

# Likelihood of the Gaussian mixture model

Likelihood of a data point $\mathbf{x}$:

$$p(\mathbf{x} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Full log-likelihood

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

No assignment variables $z_{k,n}$ for now!

# Derivation of the EM update

Let us define the responsibility as

$$\gamma(z_{k,n}) := \mathbb{E}[z_{k,n}] = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)},$$

which is the probability of $\mathbf{x}_n$ being assigned to cluster/component $k$.

Taking the derivative of the log-likelihood w.r.t. $\boldsymbol{\mu}$ and setting it to zero we recover the EM update

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{k,n}) \mathbf{x}_n,$$

with $N_k = \sum_{n=1}^{N} \gamma(z_{k,n})$.

# Mixture Models and Matrix Factorization 1/2

Assumption: $z_{k,n} \in [0,1]$ and $\sum_{k=1}^{K} z_{k,n} = 1 \;\forall n$

$$
\begin{aligned}
\|\mathbf{X} - \mathbf{U}\mathbf{Z}\|_2^2 &= \sum_{n=1}^{N} \|\mathbf{x}_n - \sum_{k=1}^{K} z_{k,n}\mathbf{u}_k\|_2^2 \\
&= \sum_{n=1}^{N} \left( \|\mathbf{x}_n\|^2 - 2\sum_{k=1}^{K} \mathbf{x}_n^T \mathbf{u}_k z_{k,n} + \left\| \sum_{k=1}^{K} z_{k,n}\mathbf{u}_k \right\|_2^2 \right) \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{k,n} \|\mathbf{x}_n - \mathbf{u}_k\|_2^2 \\
&\quad - \sum_{n=1}^{N}\sum_{k=1}^{K} z_{k,n} \left\| \mathbf{u}_k - \sum_{k'=1}^{K} z_{k',n}\mathbf{u}_{k'} \right\|_2^2
\end{aligned}
$$

We have added and subtracted the term $\sum_{n=1}^{N}\sum_{k=1}^{K} z_{k,n}\|\mathbf{u}_k\|^2$ to complete the variance term. More details in a separate document.

# Mixture Models and Matrix Factorization 2/2

Maximizing the GMM likelihood is related to the following minimization problem:

$$\min_{\mathbf{Z},\mathbf{U}} \sum_{n=1}^{N} \sum_{k=1}^{K} z_{k,n} \|\mathbf{x}_n - \mathbf{u}_k\|_2^2$$

with $z_{k,n} \in [0,1]$ and $\sum_k z_{k,n} = 1$. Which is in turn an upper bound on the problem

$$\min_{\mathbf{U},\mathbf{Z}} \|\mathbf{X} - \mathbf{U}\mathbf{Z}\|_2^2 \quad \text{with} \quad z_{k,n} \in [0,1] \text{ and } \sum_k z_{k,n} = 1.$$

However, did not present an algorithm that directly optimizes this.

# Why is sampling important?

## Normalization

Consider the Bayes rule for inference:

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{\int_{states} p(\mathbf{x}|\theta)p(\theta)d\theta}$$

To obtain the posterior $p(\theta|\mathbf{x})$ given the prior $p(\theta)$ and the likelihood $p(\mathbf{x}|\theta)$, the normalizing factor in Bayes theorem needs to be computed.

## Marginalization

Given the joint posterior, we may be interested in the marginal posterior

$$p(\theta|\mathbf{x}) = \int_{\mathcal{Z}} p(\theta, z|\mathbf{x})dz$$

# Assumptions

1. We cannot sample from $p(\mathbf{x})$
2. There is a simpler density $q(\mathbf{x})$ from which we can sample.
   - $q(\mathbf{x})$ called the sampler density.
3. We can evaluate $p^*(\mathbf{x})$ at any given point $\mathbf{x}$, which is defined as:

   $$p(\mathbf{x}) = \frac{p^*(\mathbf{x})}{Z_p}$$

   where $Z_p$ is unknown.
4. Similarly, we have:

   $$q(\mathbf{x}) = \frac{q^*(\mathbf{x})}{Z_q}$$

# Importance Sampling

Importance sampling is a method for estimating the expectation of the function $f(\mathbf{x})$.

- We sample from $q(\mathbf{x})$ instead of $p(\mathbf{x})$.
- The the importance of each point is determines by

$$w_r = \frac{p^*(\mathbf{x}^{(r)})}{q^*(\mathbf{x}^{(r)})}$$

- The expectation is thus

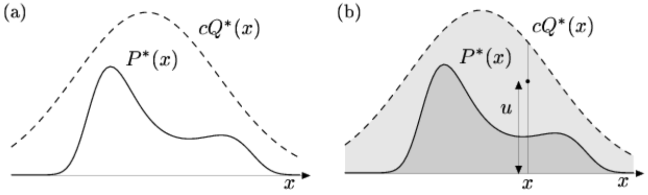
$$\hat{\mathbb{E}}(f(\mathbf{x})) = \frac{\sum_r w_r f(\mathbf{x}^{(r)})}{\sum_r w_r}$$

# Rejection Sampling

**Additional assumption**: there is a constant $c$ such that

$$cq^*(x) > p^*(x), \text{ for all } x$$

1. Generate $x$ from proposal density $q(x)$.
2. Generate a uniformly distributed random variable $u$ from the interval $[0, cq^*(x)]$.
3. If $u > p^*(x)$ then $x$ is rejected, otherwise it is added to our samples $\{x^{(r)}\}$.



# Pen&Paper

Consider Rejection sampling and answer the following questions:

1. What is the distribution over original values of $x$?
2. For the sample $x$ drawn from this distribution, what is the probability to be accepted?
3. According to the previous parts, what is the probability that a sample is accepted?
4. Discuss what happens if $c$ is chosen too small or too large.

# Assignment: Apply GMM to Color Quantization

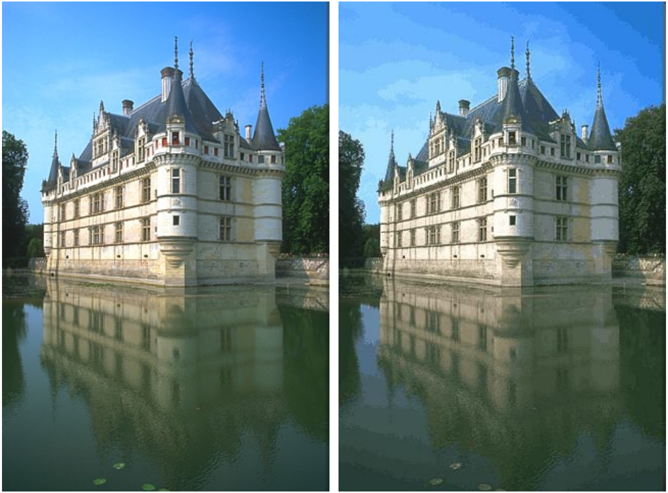


Figure: Original image (left) and compressed image (right).

# Three subparts

The assignment is very similar to the first assignment about image compression using PCA.

- **Feature extraction**: Vectorize the image. Go from $M_1 \times M_2 \times 3$ to $3 \times (M_1 \cdot M_2)$.
- **GMM clustering**: We provide you with a template for the implementation.
  - The complete E-step is implemented.
  - The only thing missing is the M-step.
- **Compress and Decompress**: Use `gmm.m` and other necessary functions for compressing and decompressing the image.