Datasta Tietoon, Autumn 2011

SOLUTIONS TO EXERCISES 2

H2 / Problem 1.
a) See the figure below left.

b) Compute first mean and substract it from $\mathbf{X}$.:

$$E\{\mathbf{x}\} = \frac{1}{4}\sum \mathbf{x}(i) = \begin{bmatrix}5\\4\end{bmatrix}$$

Thus the normalized data matrix is

$$\mathbf{X}_0 = \begin{bmatrix}-3 & 0 & 1 & 2\\ -3 & -1 & 1 & 3\end{bmatrix}$$

c) The covariance matrix is

$$\mathbf{C}_x = \frac{1}{4}\mathbf{X}_0\mathbf{X}_0^T = \frac{1}{4}\begin{bmatrix}14 & 16\\ 16 & 20\end{bmatrix}$$

The eigenvalues are computed from $\mathbf{C}_x\mathbf{u} = \lambda\mathbf{u}$, or by multiplying with 4, $\begin{bmatrix}14 & 16\\ 16 & 20\end{bmatrix}\mathbf{u} = \mu\mathbf{u}$ where μ is 4 times λ. (It may be easier to solve the equation if the coefficients are integer numbers).

We have determinant $\begin{vmatrix}14-\mu & 16\\ 16 & 20-\mu\end{vmatrix} = 0$ which gives the characteristic equation $(14-\mu)(20-\mu) - 256 = 0$ or $\mu^2 - 34\mu + 24 = 0$. The roots are 33.28 and 0.72, hence the eigenvalues λ of the covariance matrix are these divided by 4, $\lambda_1 = 8.32$ and $\lambda_2 = 0.18$.

The eigenvector $\mathbf{u}$ corresponding to the larger eigenvalue λ_1 can be computed from $\mathbf{C}_x\mathbf{u} = \lambda_1\mathbf{u}$ by

$$\begin{bmatrix}14 & 16\\ 16 & 20\end{bmatrix}\begin{bmatrix}u_1\\u_2\end{bmatrix} = 33.28\begin{bmatrix}u_1\\u_2\end{bmatrix}$$

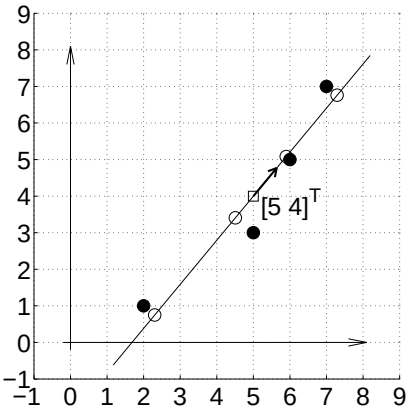
$$\begin{cases}14u_1 + 16u_2 = 33.28u_1\\ 16u_1 + 20u_2 = 33.28u_2\end{cases} \Rightarrow u_1 = 0.83u_2$$

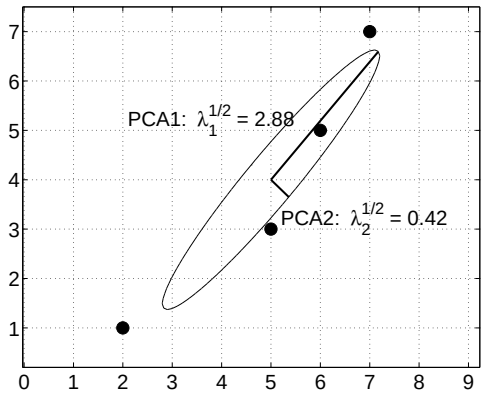
$$\Rightarrow \mathbf{u}_1 = a\begin{bmatrix}0.83\\1\end{bmatrix}, \quad a \in \mathbb{R}$$

After normalization to the unit length ($\mathbf{u}/||\mathbf{u}||$) the eigenvector corresponding largest eigenvalue λ_1 is $\mathbf{u} = [0.64\ 0.77]^T$.

The empty circles in the figure (below, left) are the projections onto 1D hyperplane (PCA1 line) by $Y = \mathbf{u}_1^T\mathbf{X}_0 = [-4.23\ -0.77\ 1.41\ 3.59]$. First PCA1 axis explains $8.32/(8.32+0.18) \approx 97.9$ % of the total variance.

In the other figure (right) it can be seen that the propotional length of the axis of an ellipse in direction PCA1 is $\sqrt{\lambda_1} \approx 2.88$, and in direction PCA2 $\sqrt{\lambda_2} \approx 0.42$. The ellipse is derived from the Gaussian covariance matrix $\mathbf{C}_x$.





In PCA the original coordinate axis system (x_1, x_2) is shifted by mean and rotated by $[\mathbf{u}_1\ \mathbf{u}_2]$. The variance of the data is maximized in direction PCA1. Data is linearly uncorrelated in the new axis system (PCA1, PCA2).

H2 / Problem 2.

We can use the *Lagrange optimization* principle for a constrained maximization problem. The principle is saying that if we need to maximize $E\{(\mathbf{w}^T\mathbf{x})^2\}$ under the constraint $\mathbf{w}^T\mathbf{w} = 1$, we should find the zeroes of the gradient of

$$E\{(\mathbf{w}^T\mathbf{x})^2\} - \lambda(\mathbf{w}^T\mathbf{w} - 1)$$

where λ is the Lagrange constant.

We can write $E\{(\mathbf{w}^T\mathbf{x})^2\} = E\{(\mathbf{w}^T\mathbf{x})(\mathbf{x}^T\mathbf{w})\} = \mathbf{w}^T E\{\mathbf{x}\mathbf{x}^T\}\mathbf{w}$ because inner product is symmetrical and the E or expectation means computing the mean over the sample $\mathbf{x}(1), ..., \mathbf{x}(n)$, thus $\mathbf{w}$ can be taken out.

We need the following general result: if $\mathbf{A}$ is a symmetrical matrix, then the gradient of the quadratic form $\mathbf{w}^T\mathbf{A}\mathbf{w}$ equals $2\mathbf{A}\mathbf{w}$. It would be very easy to prove this by taking partial derivatives with respect to the elements of $\mathbf{w}$. This is a very useful formula to remember.

Now the gradient of the Lagrangian becomes:

$$2E\{\mathbf{x}\mathbf{x}^T\}\mathbf{w} - \lambda(2\mathbf{w}) = 0$$

or

$$E\{\mathbf{x}\mathbf{x}^T\}\mathbf{w} = \lambda\mathbf{w}$$

This is the eigenvalue - eigenvector equation for matrix $E\{\mathbf{x}\mathbf{x}^T\}$. But there are d eigenvalues and vectors: which one should be chosen?

Multiplying from the left by $\mathbf{w}^T$ and remembering that $\mathbf{w}^T\mathbf{w} = 1$ gives

$$\mathbf{w}^T E\{\mathbf{x}\mathbf{x}^T\}\mathbf{w} = \lambda$$

showing that λ should be chosen as the largest eigenvalue in order to maximize $\mathbf{w}^T E\{\mathbf{x}\mathbf{x}^T\}\mathbf{w} = E\{y^2\}$. This was to be shown.

H2 / Problem 3.

Problem: Given data sample $\mathbf{X}$, compute estimator $\hat{\lambda}_{\mathrm{ML}}$ with which data has been most probably generated. The maximum likelihood (ML) method is summarized as (p. 233. Milton, Arnold: Introduction to probability and statistics. Third edition. McGraw-Hill, 1995)

1. Obtain a random sample $\mathbf{X} = \{x(1), x(2), \ldots, x(n)\}$ from the distribution of a random variable X with density p and associated parameter θ
2. Define a likelihood function $L(\theta)$ by
$$L(\theta) = \prod_{i=1}^{n} p(x(i))$$
3. Find the expression for θ that maximizes the likelihood function. This can be done directly or by maximizing $\ln L(\theta)$
4. Replace θ by $\hat{\theta}$ to obtain an expression for ML estimator for θ
5. Find the observed value of this estimator for a given sample

Let us assume that data samples $\mathbf{X} = \{x(1), x(2), \ldots, x(n)\}$ are i.i.d., that is, they are independent and identically-distributed. Independence means that joint density function $P(A, B, C)$ can be decomposed to product of marginal density functions: $P(A, B, C) = P(A) \cdot P(B) \cdot P(C)$. Each sample $x(i)$ is from the same (identical) distribution $p(x|\lambda)$ with the same λ. One-dimensional exponential propability density function is $p(x|\lambda) = \lambda e^{-\lambda x}$ where rate $\lambda = 1/\mu$.

In this case the likelihood function $L(\lambda)$ ("uskottavuusfunktio") is

$$L(\lambda) = p(\mathbf{X}|\lambda) = p(x(1), x(2), \ldots, x(n)|\lambda) \overset{\text{i.i.d.}}{=} \prod_{i=1}^{n} p(x(i)|\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x(i)}$$

Because samples $x(i)$ are known, likelihood is a function of λ only.

In order to find estimator $\hat{\lambda}_{\mathrm{ML}}$ we maximize likelihood by, e.g., setting derivative to zero. Because we are finding the extreme point, we can take logarithm and still find the same maximum point of likelihood function. The computation comes much easier because $\ln(A \cdot B \cdot C) = \ln A + \ln B + \ln C$. The log-likelihood:

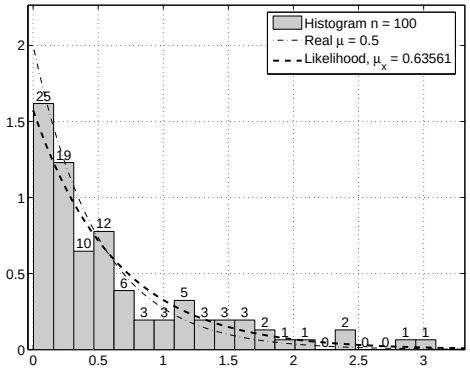
$$\begin{aligned} \ln L(\lambda) = \ln p(\mathbf{X}|\lambda) &= \ln \prod_{i=1}^{n} [\lambda e^{-\lambda x(i)}] \\ &= \sum_{i=1}^{n} [\ln \lambda - \lambda x(i)] \\ &= n \ln \lambda - \lambda \sum_{i=1}^{n} x(i) \end{aligned}$$

Putting the derivative with respect to λ to zero gives the solution $\hat{\lambda}_{\mathrm{ML}}$

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\lambda} \ln L(\lambda) = \frac{\mathrm{d}}{\mathrm{d}\lambda} \Big\{ n \ln \lambda - \lambda \sum_{i=1}^{n} x(i) \Big\} &= \frac{n}{\lambda} - \sum_{i=1}^{n} x(i) = 0 \\ \frac{1}{\lambda} &= \frac{1}{n} \sum_{i=1}^{n} x(i) \end{aligned}$$

Thus the ML (maximum likelihood) estimate for λ is the inverse of the mean value of the sample.

An example in the figure below. $n = 100$ samples $x(1), \ldots, x(100)$ are drawn from the exponential distribution with $\mu = 0.5 \Leftrightarrow \lambda = 1/\mu = 2$. Sample mean $1/\hat{\lambda}_{\mathrm{ML}} = \bar{X} = 0.636$ at this time.

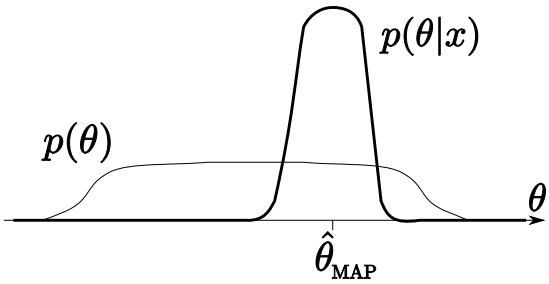


H2 / Problem 4.

Problem: Given data sample $\mathbf{x}$ and prior distribution for μ, compute estimator $\hat{\mu}_{\mathrm{MAP}}$. This Bayesian maximum posterior (MAP) method follows that of maximum likelihood (Problem H2/3) but now the function to be maximized is not likelihood but posterior = likelihood $\times$ prior. Inference using Bayes' theorem can be written as

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

where $p(\theta|x)$ is posterior, $p(x|\theta)$ likelihood, $p(\theta)$ prior and $p(x)$ evidence, which is just a scaling factor. θ contains all parameters. This results to a posterior distribution (our knowledge after seeing data) with respect to θ which is more exact (with smaller variance) than prior (our knowledge or guess before seeing any data), see figure below. Note that here we have a distribution for θ whereas the maximum likelihood gives a point estimate. Finally, however, the MAP estimate is a single value that maximizes the posterior.



Let us again assume that data samples $\mathbf{X} = \{x(1), x(2), \ldots, x(n)\}$ are i.i.d., that is, they are independent and identically-distributed. The one-dimensional normal (Gaussian) density function is

$$p(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Each sample $x(i)$ is from the same (identical) distribution $p(x|\mu, \sigma)$ with the same μ and σ.

Likelihood function is

$$L(\mu, \sigma) = p(\mathbf{X}|\mu, \sigma) \overset{\text{i.i.d.}}{=} \prod_{i=1}^{n} p(x(i)|\mu, \sigma)$$

Our prior for μ (with hyperparameters $\mu_0 = 0$ and $\sigma_0 = 1$) is

$$p(\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2}}$$

Posterior is

$$p(\mu, \sigma|\mathbf{X}) \propto L(\mu, \sigma)p(\mu)p(\sigma)$$

where the constant denominator $p(\mathbf{X})$ can be omitted when searching maximum. The symbol $\propto$ can be read "is propotional".

Taking logarithm of likelihood function and setting the derivative with respect to μ to zero follows computation as in Problem H2/3. The log-likelihood:

$$\begin{aligned}
\ln L(\mu, \sigma) = \ln p(\mathbf{X}|\mu, \sigma) &= \ln \prod_{i=1}^{n} [\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x(i)-\mu)^2}{2\sigma^2}}] \\
&= \sum_{i=1}^{n} [\ln(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x(i)-\mu)^2}{2\sigma^2}})] \\
&= \sum_{i=1}^{n} [-\ln\sigma - \ln(\sqrt{2\pi}) - \frac{(x(i)-\mu)^2}{2\sigma^2}]
\end{aligned}$$

The log-prior probability for μ is

$$\ln p(\mu) = -\ln(\sqrt{2\pi}) - \frac{1}{2}\mu^2$$

The log-posterior can be written with Bayes' theorem as a sum of log-likelihood and log-prior

$$\ln p(\mu, \sigma|\mathbf{X}) \propto \ln L(\mu, \sigma) + \ln p(\mu) + \ln p(\sigma)$$

In other words, all parts depending on μ in the Bayesian log-posterior probability are:

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[(x(i)-\mu)^2]-\frac{1}{2}\mu^2$$

Setting derivative of log-posterior with respect to μ to zero gives

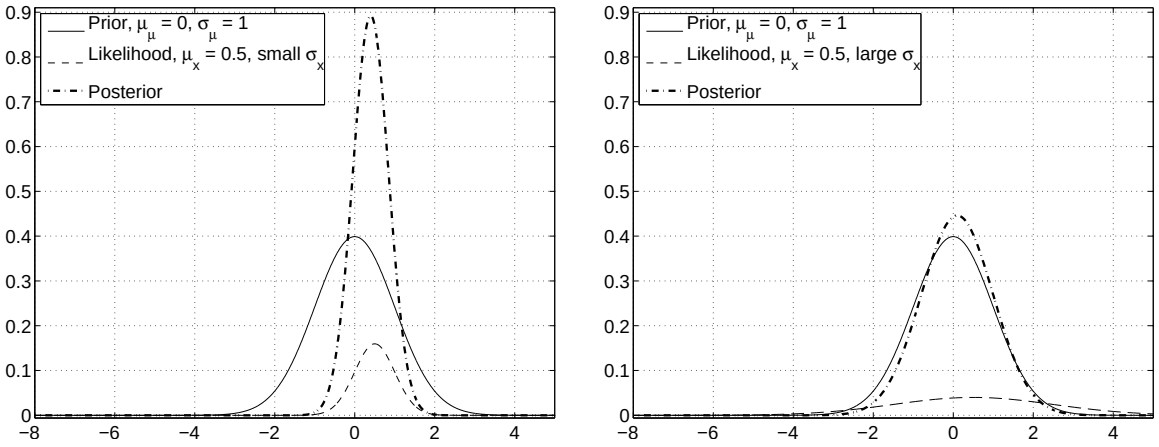
$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}\mu}\Big\{\Big(-\frac{1}{2\sigma^2}\Big)\sum_{i=1}^{n}[(x(i)-\mu)^2]-\frac{1}{2}\mu^2\Big\} \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}[2(x(i)-\mu)(-1)]-\mu \\ &= \sum_{i=1}^{n}[x(i)]-n\mu-\sigma^2\mu \end{aligned}$$

which finally gives $\hat{\mu}_{\mathrm{MAP}}$

$$\mu=\frac{1}{n+\sigma^2}\sum_{i=1}^{n}x(i)$$

The interpretation is as follows: if the variance σ^2 of the sample is very small, then the sample can be trusted. Therefore μ is very close to the sample mean $\frac{1}{n}\sum_{i=1}^{n}x(i)$ (likelihood estimate). See an example in the figure below left: $\hat{\mu}_{\mathrm{MAP}}\approx 0.48$ (posterior) is close to $\mu_{\mathrm{ML}}=0.5$ (likelihood).

On the other hand, if σ^2 is very large, then the sample cannot be trusted and the prior information dominates. Density function of μ becomes close to that of prior assumption. See an example in the figure below right: $\hat{\mu}_{\mathrm{MAP}}\approx 0.04$ (posterior) is close to $\mu_{\mathrm{PRIOR}}=0$.



In case of maximum likelihood, the estimator is $\hat{\mu}_{\mathrm{ML}}=\frac{1}{n}\sum_{i=1}^{n}x(i)=\bar{X}$. The only, but remarkable difference is the variance term in the denominator.