UML2014
Exercise set 3
Solutions to be presented in the 28.3.2014 session

Exercise 1:

1.1 Show that

$$(A + \varepsilon B)^{-1} = A^{-1} - \varepsilon A^{-1} B A^{-1} + O(\varepsilon^2) \tag{1}$$

in the limit $\varepsilon \to 0$

1.2 Show that

$$\frac{\partial}{\partial A_{ij}} (A^{-1})_{nm} = -(A^{-1})_{ni} (A^{-1})_{jm} \tag{2}$$

1.3 Assume that $f : \mathbb{R}^{N \times N} \to \mathbb{R}$ is some differentiable function whose partial derivatives are known. Find some useful formula for

$$\frac{\partial g(A)}{\partial A_{ij}} \tag{3}$$

when g has been defined with the formula $g(A) = f(A^{-1})$.

Advice: We can assume that the Neumann series

$$(\mathrm{id} + A)^{-1} = \mathrm{id} - A + A^2 - A^3 + \cdots \tag{4}$$

is already known. (You can study this series too, though, if you are not familiar with it...) In 1.1 you must begin with $A + \varepsilon B = A(\mathrm{id} + \varepsilon A^{-1} B)$ or $A + \varepsilon B = (\mathrm{id} + \varepsilon B A^{-1}) A$. In 1.3 you must use the chain rule of differentiation. It is simple with vectors, but the matrix form of the input parameter can make it feel more complicated. One possibility is to define a function $I : \mathbb{R}^{N \times N} \to \mathbb{R}^{N \times N}$, $I(A) = A^{-1}$, and see g as $g = f \circ I$.

Exercise 2:

In this exercise, we calculate the gradient of

$$J(W) = \log |\det(W)|. \tag{5}$$

2.1 Assume that $\mathbf{u}_1, \ldots, \mathbf{u}_N$ are linearly independent eigenvectors of W, and form a matrix $U = (\mathbf{u}_1, \ldots, \mathbf{u}_N)$. Let's then define $V = (U^{-1})^T$, and define vectors $\mathbf{v}_1, \ldots, \mathbf{v}_N$ so that $V = (\mathbf{v}_1, \ldots, \mathbf{v}_N)$. Show that the eigenvalues λ_n can be written by formula

$$\lambda_n = \mathbf{v}_n^T W \mathbf{u}_n \tag{6}$$

2.2 We can consider λ_n, $\mathbf{u}_n$ and $\mathbf{v}_n$ to be functions of W, so that they can be written as $\lambda_n(W)$, $\mathbf{u}_n(W)$ and $\mathbf{v}_n(W)$. Show that $\nabla \lambda_n(W) = \mathbf{v}_n \mathbf{u}_n^T$ by first substituting $\lambda_n(W) = \mathbf{v}_n(W)^T W \mathbf{u}_n(W)$. You can assume that $\mathbf{u}_n(W)$ and $\mathbf{v}_n(W)$ are differentiable functions of W.

2.3 Use the formula for $\nabla\lambda_n(W)$ to show that $\nabla J(W) = (W^{-1})^T$.

Advice: In 2.2 it is probably clearer to begin searching for $\frac{\partial}{\partial W_{ij}}\lambda_n(W)$ with some fixed i, j. At some point you must use the fact that $\mathbf{u}_n$ is a right eigenvector of W, and $\mathbf{v}_n$ a left eigenvector of W, so seek opportunities for these if you seem to get stuck. In 2.3 you must recall last week's exercises.

Exercise 3:

A Gaussian random vector $\mathbf{x}$ of dimension N has the density

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{N/2}\det(\Sigma)^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right], \tag{7}$$

where Σ is the covariance matrix and $\boldsymbol{\mu}$ is the mean.

3.1 Given independently distributed data $\mathbf{x}_1, \dots, \mathbf{x}_K$ where each sample $\mathbf{x}_k$ is a Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix Σ, formulate the log-likelihood $\ell(\boldsymbol{\mu}, \Sigma)$.

3.2 Calculate the gradient of $\ell(\boldsymbol{\mu}, \Sigma)$ with respect to $\boldsymbol{\mu}$ and Σ.

3.3 Conclude that the ML (maximum likelihood) estimate for $\boldsymbol{\mu}$ is the sample mean

$$\hat{\boldsymbol{\mu}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_k, \tag{8}$$

and that the ML estimate for Σ is the sample covariance matrix

$$\hat{\Sigma} = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{x}_k-\hat{\boldsymbol{\mu}})(\mathbf{x}_k-\hat{\boldsymbol{\mu}})^T, \tag{9}$$

You can omit the study of Hessian for simplicity.

Exercise 4:

Assume that $X \in \mathbb{R}^{N\times K}$ is a matrix of K sample vectors. We denote the sample vectors as X_{*k}, which with fixed k are $N \times 1$ matrices. Let $\mathbf{w} \in \mathbb{R}^{N\times 1}$ be a vector which reduces the dimension of the data from N to 1 via $Z = \mathbf{w}^T X$. So Z is a $1 \times K$ matrix. Next we want to reconstruct X from Z via

$$\hat{X} = \mathbf{w}Z. \tag{10}$$

Notice that $\hat{X}$ is again a $N \times K$ matrix. In this exercise, we will show that taking for $\mathbf{w}$ the first principal component is the optimal way of reducing the dimension, if the optimality criterion is the squared reconstruction error J

$$J(\mathbf{w}) = \frac{1}{K}\sum_{k=1}^{K}\left\|X_{*k}-\hat{X}_{*k}\right\|^2 \tag{11}$$

Here Z and $\hat{X}$ are considered to be functions of $\mathbf{w}$, so that we could denote $Z = Z(\mathbf{w})$ and $\hat{X} = \hat{X}(\mathbf{w})$.

4.1 Prove that minimizing $J(\mathbf{w})$ with constraint $\|\mathbf{w}\| = 1$ is equivalent to maximizing $\mathbf{w}^T X X^T \mathbf{w}$ (with the same constraint).

4.2 Assume that $\lambda_1 > \cdots > \lambda_N$ are some fixed numbers, and that parameters $m_1, \ldots, m_N$ must satisfy $m_n \geq 0$ and $m_1 + \cdots + m_N = 1$. Prove, that under these constraints, the quantity

$$\lambda_1 m_1 + \cdots + \lambda_N m_N \tag{12}$$

is maximized by choosing $(m_1, m_2, \ldots, m_N) = (1, 0, \ldots, 0)$.

4.3 Assume that A is some real symmetric matrix with distinct eigenvalues. Prove that if the vector $\mathbf{w}$ must satisfy $\|\mathbf{w}\| = 1$, the quantity $\mathbf{w}^T A \mathbf{w}$ will be maximized with respect to $\mathbf{w}$ precisely when $\mathbf{w}$ is an eigenvector of A corresponding to the largest eigenvalue.

4.4 Deduce that minimizing $J(\mathbf{w})$ with constraint $\|\mathbf{w}\| = 1$ is equivalent to finding an eigenvector of XX^T with the largest eigenvalue.

Advice: In 4.1 one possibility is to use the definition of the norm $\|\mathbf{x}\|^2 = \sum_{n=1}^{N} x_n^2$ and see what happens. This will lead to plenty of work, but the path will not be impossible. A smarter alternative is to notice that the objective function can be written in the form $J(\mathbf{w}) = \frac{1}{K}\mathrm{Tr}\big((X^T - \hat{X}^T)(X - \hat{X})\big)$. After substituing $\hat{X} = \mathbf{w}\mathbf{w}^T X$, a nicer path to the end result might begin to appear. The claim in 4.2 can be proven in several different ways. One possibility is to use $m_1 = 1 - m_2 - \cdots - m_N$ to transform the N dimensional constrained optimization problem into an $N-1$ dimensional unconstrained optimization problem (or differently constrained). An alternative way is to prove that if $(m_1, m_2, \ldots, m_N) \neq (1, 0, \ldots, 0)$ (anti-thesis), then

$$\lambda_1 > \lambda_1 m_1 + \cdots + \lambda_N m_N. \tag{13}$$

This claim is equivalent to

$$\lambda_1 > \lambda_2 \frac{m_2}{1 - m_1} + \cdots + \lambda_N \frac{m_N}{1 - m_1}. \tag{14}$$