ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Computational Intelligence Lab

2012

Prof. J. M. Buhmann

Final Exam (Repetition)

28 January 2013

First and Last name: _____

ETH number: _____

Signature: _____

General Remarks

- Please check that you have all 17 pages of this exam.
- Remove all material from your desk which is not permitted by the examination regulations.
- Fill in your name and ETH number and sign the exam. Place your student ID on the desk.
- You have 120 minutes for the exam. There are four questions, where you can earn a total of 100 points.
- Write your answers directly on the exam sheets. If you need more space, put your name and ETH number on top of each supplementary sheet.
- Answer the questions in English. Please use a black or blue pen to answer the questions.
- You may provide at most one valid answer per question. Invalid solutions must be canceled out clearly.

	Topic	Max. Points	Points Achieved	Visum
1	Dimensionality Reduction	25		
2	Clustering	25		
3	Sparse Coding	25		
4	Robust PCA	25		
Total		100		

Grade: .....

Question 1: Dimensionality Reduction (25 pts.)

a) Cross **all** of the correct statements.

- ☐ In PCA using the largest eigenvalues ensures that the reconstruction has zero error.
- ☐ Performing SVD on a symmetric matrix has the same outcome as PCA.
- ☐ In PCA the projected components are uncorrelated and therefore the reconstruction has zero error.
- ☐ The model selection problem exists for SVD but not for PCA.
- ☐ In PCA the objective is to minimize the low dimension approximation error under the L_2 distance.

3 pts

b) Let $\mathbf{A}$ be the following matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 4 \\ 2 & 0 \\ 0 & -3 \end{bmatrix}$$

We would like to find its singular value decomposition, $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$

.

1. Show that the matrices $\mathbf{V}$ and $\mathbf{D}$ (having a decreasing order of the singular values) are as follows:

$$\mathbf{V} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} 5 & 0 \\ 0 & 2 \\ 0 & 0 \end{bmatrix}$$

5 pts

2. Complete the SVD by computing $\mathbf{U}$. Verify that $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$.

4 pts

c) Recall the image compression application with PCA we saw in class. Suppose that the images we would like to compress are represented by a matrix $\Im$ of values in $[0, 1]$ of size 100×150. Each entry value takes a byte to store. We perform feature extraction by considering non-overlapping patches of some size, each patch is then vectorized into a column vector in the data matrix $\mathbf{X}$. The next step is to perform PCA on $\mathbf{X}$.

1. If we would use patches of size 10×15, What would be the dimensions of the covariance matrix?

2 pts

2. Give a detailed description of a compression scheme using PCA, under the requirement that the resulting image can take up to $5,000$ bytes.

4 pts

d) Let $\mathbf{A}$ be an $N \times N$ symmetric matrix, we say that $\mathbf{A}$ is *positive semi-definite* (PSD) if for all vectors
$\mathbf{w} \in \mathbb{R}^N$ the following holds

$$\mathbf{w}^\top \mathbf{A} \mathbf{w} \geq 0$$

Prove that if all of the eigenvalues of $\mathbf{A}$ are greater or equal to zero, i.e. $\lambda_i \geq 0$ for all eigenvalues λ_i of $\mathbf{A}$, then $\mathbf{A}$ is PSD.
Hints:

1. Any vector $\mathbf{w} \in \mathbb{R}^N$ can be re-represented using any basis of $\mathbb{R}^N$.
2. $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$.

7 pts

Question 2: Clustering (25 pts.)

a) For each of the following statements, determine whether it is true or false.

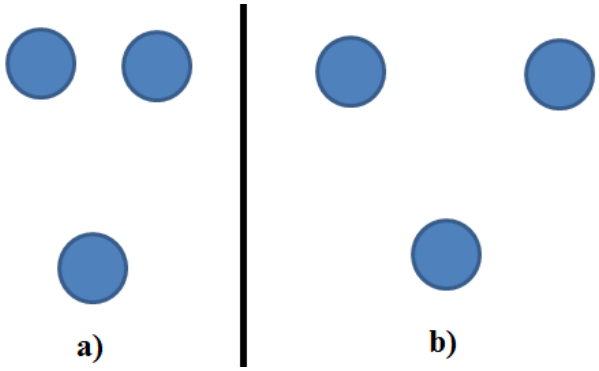
1. The Expectation-Maximization (EM) algorithm is faster than K-means algorithm.
   True/False

2. Gaussian Mixture Model (GMM) clustering is equivalent to using soft assignments instead of K-means hard assignments.
   True/False

3. The Bayesian Information Criterion (BIC) score cannot be smaller than the Akaike Information Criterion (AIC) score for large enough datasets and fixed number of free parameters.
   True/False

3 pts

b) Recall the notion of clustering stability. In which of the following situations the stability measure might fail to find the correct number of clusters (the true number of clusters is 3)? Explain your answer.



2 pts

c) We use a mixture of K component distributions to model the frequencies of $m = 100$ different terms in a collection of $n = 2000$ independent documents. The log-likelihood function of the n documents is written as:

$$p(\mathbf{D}|\boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \log \sum_{c=1}^{K} \pi_c f(D_i|\boldsymbol{\mu}_c), \tag{1}$$

The component distributions $f(D_i|\boldsymbol{\mu}_c)$ are defined as:

$$f(D_i|\boldsymbol{\mu}_c) = \prod_{j=1}^{m} \mu_{jc}^{F_{ij}}. \tag{2}$$

where μ_{jc} is the probability of the j^{th} term in the c^{th} component distribution and F_{ij} is the frequency of the j^{th} term in document D_i.

1. Introduce the suitable latent variables that you need for maximizing the log-likelihood function. **2 pts**

2. Calculate the expectation of the the latent variables. **3 pts**

3. Use the expectation of the latent variables to calculate the unknown parameters μ_{jc}. Write down the details of your calculations.
   Hint: Note that $\sum_{j=1}^{m} \mu_{jc} = 1, \forall 1 \leq c \leq K$. **5 pts**

4. Suppose we already know that $\pi_1 = \pi_2$. We use Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to determine the number of component distributions. For the given K and log-likelihood numbers compute the AIC and BIC scores. Answer:

K	log-likelihood	AIC	BIC
3	-4,000	.....	.....
5	-3,500	.....	.....

4 pts

d) In the Role-Based Access Control (RBAC) problem, we are given the user-permission matrix $\mathbf{X}$ and the goal is to find roles $\mathbf{U}$ and assignments $\mathbf{Z}$ such that $\mathbf{X} = \mathbf{U} \otimes \mathbf{Z} \iff x_{dn} = \bigvee_k [u_{dk} \wedge z_{kn}]$, with $\mathbf{X} \in \{0,1\}^{D \times N}$, $\mathbf{U} \in \{0,1\}^{D \times K}$ and $\mathbf{Z} \in \{0,1\}^{K \times N}$. Moreover suppose that $\boldsymbol{\beta} = (p\{u_{dk} = 0\})^{D \times K}$.

1. We focus on the multi-assignment clustering (MAC) model. Consider one entry x_{dn} of matrix $\mathbf{X}$. Compute $p(x_{dn} = 1 | \boldsymbol{\beta}_{\boldsymbol{d}\cdot}, \mathbf{z}_{\cdot \mathbf{n}})$.

3 pts

2. We consider a mixture noise generation model where ξ_{dn} indicates whether a bit is generated by noise (i.e. $\xi_{dn} = 0$) or by $(\mathbf{U} \otimes \mathbf{Z})_{dn}$ (i.e. $\xi_{dn} = 1$). The noisy bit is drawn from a Bernoulli distribution: $p_N(x_{dn}|r) = r^{x_{dn}}(1-r)^{1-x_{dn}}$. Complete the following statement (note that $p_S(x_{dn}|\boldsymbol{\beta}_{\boldsymbol{d}\cdot}, \mathbf{z}_{\cdot\mathbf{n}})$ shows the distribution of x_{dn} if generated by the roles).

$$p\left(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\xi}, r\right) = \prod_{n,d} r^{\cdots\cdots} \; (1-r)^{\cdots\cdots} \; p_S(x_{dn}|\boldsymbol{\beta}_{\boldsymbol{d}\cdot}, \mathbf{z}_{\cdot\mathbf{n}})^{\cdots\cdots}$$

3 pts

Question 3: Sparse Coding (25 pts.)

Orthogonal Matrices and Bases

a) What kind of computational issue can be avoided by using an *orthonormal* bases? Briefly explain your answer.

2 pts

b) Check **all** the correct statements.
If $\mathbf{U} \in \mathbb{R}^{D \times D}$ is an *orthogonal* matrix and $\mathbf{x}$,$\mathbf{y} \in \mathbb{R}^{D}$ then:

- ☐ $\operatorname{rank}(\mathbf{U}) = D$
- ☐ $\det(\mathbf{U}) = 0$
- ☐ $\left\|\mathbf{U}^\top \mathbf{x}\right\|_2^2 = \|\mathbf{x}\|_2^2$
- ☐ $\mathbf{U}^\top \mathbf{x} + \mathbf{U}^\top \mathbf{y} = \mathbf{x} + \mathbf{y}$
- ☐ $\mathbf{U}^\top \mathbf{U} = \mathbb{I}$

3 pts

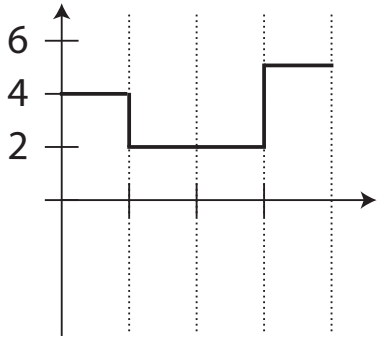
c) Check **all** of the signals below, which you would assume to be sparse in some basis.

- ☐ Recorded birds singing
- ☐ Picture of a face
- ☐ Sequence of dice throwing results
- ☐ Daily temeperature records over the year
- ☐ Sequence of directions (left or right) of a tennis player shots

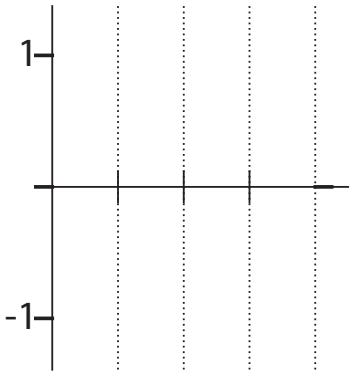
3 pts.

d) We would like to compress the signal below ($\mathbf{x} \in \mathbb{R}^4$) using Haar Wavelets. Due to the limited budget we keep only three basis functions out of 4. Draw the **remaining** basis function (corresponding to the zero valued coefficient after the *thresholding* step).

Hint: You do not have to compute the coefficients.

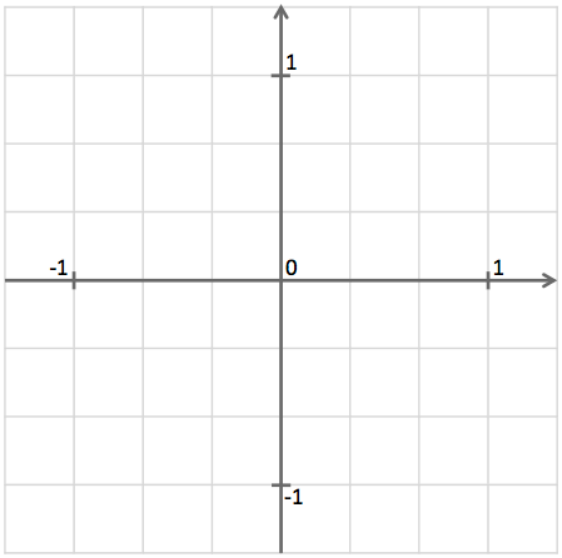


Answer: **6 pts**

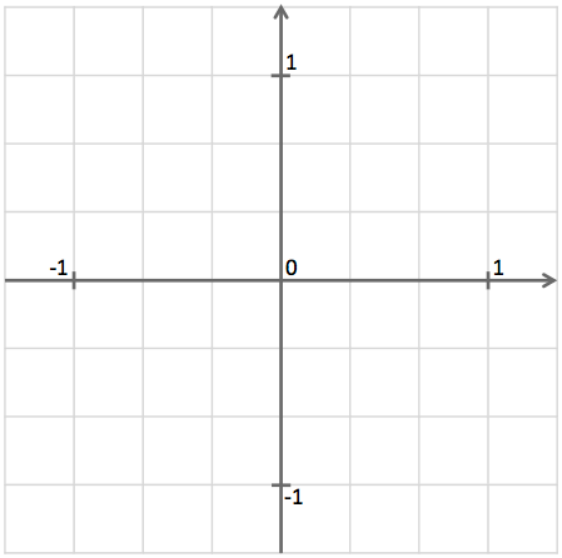


e) Assume that you have two overcomplete dictionaries $\mathbf{B}_1$ and $\mathbf{B}_2$ in a two dimensional space ($D = 2$) with $L = 4$ normalized atoms each. $\mathbf{B}_1$ was constructed by adding two atoms consequently to an orthonormal basis such that the coherence coh($\mathbf{B}_1$) was minimized at each step. $\mathbf{B}_2$ is constructed as an overcomplete dictionary with 4 atoms with the minimal coherence coh($\mathbf{B}_2$).

1. Is coh($\mathbf{B}_1$) = coh($\mathbf{B}_2$)?
2. Draw possible $\mathbf{B}_1$ and $\mathbf{B}_2$.



(1) $\mathbf{B}_1$



(2) $\mathbf{B}_2$

3. coh($\mathbf{B}_1$) =
4. coh($\mathbf{B}_2$) =
5. Prove that ($\mathbf{B}_1$) and ($\mathbf{B}_2$) are the ones which minimizes the coherence given their construction.

Answer:

11 pts

Question 4: Convex Optimization / Robust PCA (25 pts.)

a) Suppose you are given a video of Polyterasse filmed from the Dozentenfoyer. Your task is to separate foreground from background using Robust PCA.

1. How do you need to set-up the matrix $\mathbf{X}$, that you want to use as an input to the algorithm? In what matrix does the foreground end up and in what matrix the background and why?
   Answer:

2. What conditions on the foreground and background of the video need to be met for Robust PCA to work? What conditions would you need on the underlaying matrices that you want to separate, to get theoretical recovery guarantees?
   Answer:

6 pts

b) Consider the function $f(\mathbf{x}) = \max_i(x_i - b)$ from $\mathbb{R}^n \to \mathbb{R}$ with $b \in \mathbb{R}$.

1. Show that $f(\mathbf{x})$ is a convex or non-convex function by using the definition of a convex function.

Answer:

5 pts

c) Assume you are given a user-movie-rating matrix $\mathbf{X}$ in a collaborative filtering task and you want to do matrix completion to predict the missing ratings. Unfortunately some of the ratings are corrupted.

1. Write down how you can adapt Robust PCA to solve this problem (i.e. write down the optimization problem).
2. How are the constraints different from the original Robust PCA problem?
3. Explain what the variables mean in your formulation.

Answer:

6 pts

d) Find the lagrange dual problem of the following linear program in inequality form

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \mathbf{c}^\top \mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b} \end{aligned} \tag{3}$$

1. Form the Lagrangian $L(\mathbf{x}, \boldsymbol{\lambda})$ and write down the dual function $g(\boldsymbol{\lambda})$.
   Answer:

2. Find an analytical expression of the dual function $g(\boldsymbol{\lambda})$ by minimizing over $\mathbf{x}$. Use the fact, that a linear function is bounded below only when it is identically zero.
   Answer:

3. Write down the lagrange dual problem to the primal problem (3). Use the dual function from above. Your objective function should again be linear.
   Answer:

8 pts