CS5487 Problem Set 7
Linear Dimensionality Reduction

Antoni Chan
Department of Computer Science
City University of Hong Kong

Dimensionality

Problem 7.1 Shell of a hypersphere

Consider a hypersphere S_1 of radius r in $\mathbb{R}^d$, and its shell of thickness ϵ. The shell can be defined as the region between S_1 and another hypersphere S_2 with radius $r-\epsilon$. Hence, the volume of the shell is

$$V(shell) = V(S_1) - V(S_2) = \left[1 - \frac{V(S_2)}{V(S_1)}\right] V(S_1), \tag{7.1}$$

where the volume of a d-dimensional hypersphere of radius r is

$$V(S) = \frac{\pi^{d/2} r^d}{\Gamma(\frac{d}{2}+1)}. \tag{7.2}$$

(a) Show that the ratio between the volumes of the two sphere is

$$\frac{V(S_2)}{V(S_1)} = \left(1 - \frac{\epsilon}{r}\right)^d, \tag{7.3}$$

and hence, for any $0 < \epsilon < r$,

$$\lim_{d\to\infty} \frac{V(S_2)}{V(S_1)} = 0. \tag{7.4}$$

This shows that for high dimension $V(shell) = V(S_1)$. In other words, *all the volume of the hypersphere is in its shell!*

.....

Principal Component Analysis

Problem 7.2 PCA as minimizing reconstruction error

In this problem, we will prove that PCA finds a representation that minimizes the reconstruction error of the data points. Let $X = \{x_1, \cdots, x_n\}$ be the data points, with $x_i \in \mathbb{R}^d$. Let the new representation (the PCA coefficients) be $Z = \{z_1, \cdots, z_n\}$, with $z_i \in \mathbb{R}^k$. Denote $\Phi = [\phi_1, \cdots, \phi_k] \in \mathbb{R}^{d\times k}$ as a matrix of basis vectors ϕ_j, and $c \in \mathbb{R}^d$ as the offset. The basis vectors are constrained to be orthonormal, i.e., $\phi_j^T \phi_j = 1$ and $\phi_j^T \phi_i = 0$ for $i \neq j$, or succinctly $\Phi^T\Phi = I$. The reconstruction of a data point from its lower-dimensional representation is

$$\hat{x}_i = c + \Phi z_i = c + \sum_{j=1}^{k} \phi_j z_{ij}, \tag{7.5}$$

where z_{ij} is the jth element of z_i.

The total reconstruction error is given by

$$J(Z, \Phi, c) = \sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2 . \tag{7.6}$$

Hence, the goal is to find the optimal coefficients Z, basis matrix Φ, and offset c that minimize J,

$$\{Z^*, \Phi^*, c^*\} = \underset{Z,\Phi,c}{\operatorname{argmin}} J(Z, \Phi, c) \quad \text{s.t. } \Phi^T\Phi = I. \tag{7.7}$$

(a) Note that only Z depends on X, hence we can first solve for the optimal Z as a function of $\{c, \Phi\}$. Show that the optimal coefficients Z that minimize J are

$$z_i^* = \Phi^T(x_i - c). \tag{7.8}$$

(b) Substitute (7.8) into J to obtain a new objective function $\hat{J}(c, \Phi)$,

$$\hat{J}(c, \Phi) = \sum_{i=1}^{n} (x_i - c)^T (I - \Phi\Phi^T)(x_i - c). \tag{7.9}$$

(c) Show that the offset c that minimizes $\hat{J}$ is

$$c^* = \frac{1}{n} \sum_{i=1}^{N} x_i, \tag{7.10}$$

i.e., the sample mean.

(d) To find the optimal Φ^*, first show that the optimization of $\hat{J}$ can be rewritten as

$$\Phi_* = \underset{\Phi}{\operatorname{argmin}}\, \hat{J}(c^*, \Phi) \quad \text{s.t. } \Phi^T\Phi = I \tag{7.11}$$

$$= \underset{\Phi}{\operatorname{argmax}}\, \operatorname{tr}(\Phi^T\Sigma\Phi) \quad \text{s.t. } \Phi^T\Phi = I \tag{7.12}$$

$$= \underset{\Phi}{\operatorname{argmax}} \sum_{j=1}^{k} \phi_i^T \Sigma \phi_i \quad \text{s.t. } \Phi^T\Phi = I, \tag{7.13}$$

where $\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - c^*)(x_i - c^*)^T$ is the sample covariance matrix.

(e) Use Langrange multipliers to show that the stationary point for ϕ_j of the constrained optimization problem in (7.13) is an eigenvector of Σ, i.e.

$$\Sigma\phi_j = \lambda_j\phi_j, \tag{7.14}$$

where λ_j is the corresponding eigenvalue. Finally, using this fact and (7.12), show that

$$\Phi_* = \underset{\Phi}{\operatorname{argmax}}\, \operatorname{tr}(\Lambda) = \operatorname{argmax} \sum_{j=1}^{k} \lambda_i, \tag{7.15}$$

which shows that the optimal Φ_* are the eigenvectors corresponding to the k largest eigenvalues of Σ.

.....

Problem 7.3 PCA as maximizing variance

In this problem, we will show that PCA maximizes the variance of the projected data. Let $X = \{x_1, \cdots, x_n\}$ be the data points, with $x_i \in \mathbb{R}^d$. Define the set of projection directions as $\Phi = [\phi_1, \cdots, \phi_k] \in \mathbb{R}^{d \times k}$, where the ϕ_i are orthonormal, i.e., $\phi_j^T \phi_j = 1$ and $\phi_j^T \phi_i = 0$ for $i \neq j$, or succinctly $\Phi^T \Phi = I$. Let $\bar{x} = \frac{1}{n} \sum_i x_i$ be the sample mean of X.

The projected data is $Z = \{z_1, \cdots, z_n\}$, with $z_i \in \mathbb{R}^k$, where

$$z_i = \Phi^T (x_i - \bar{x}). \tag{7.16}$$

The goal is to maximize the variance of the projected data Z by selecting the optimal Φ,

$$\Phi^* = \underset{\Phi}{\operatorname{argmax}} \operatorname{tr} \left(\frac{1}{n} \sum_{i=1}^{n} (z_i - \bar{z})(z_i - \bar{z})^T \right) \quad \text{s.t. } \Phi^T \Phi = I, \tag{7.17}$$

where $\bar{z} = \frac{1}{n} \sum_i z_i$ is the sample mean of Z.

(a) Show that $\bar{z} = 0$.

(b) Use (a) to show that (7.17) is equivalent to (7.12). Hence, maximizing the variance of the projected data is equivalent to minimizing the reconstruction error, leading to the PCA formulation.

.....

Problem 7.4 PCA implementation using SVD

In this problem, we will consider implementing PCA using the *singular value decomposition* (SVD, see Problem 1.17). Let $X = [x_1, \cdots, x_n]$ be the matrix of data points with sample mean and covariance,

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \Sigma = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T. \tag{7.18}$$

Let $\bar{X}$ be the matrix of mean-subtracted points,

$$\bar{X} = [x_1 - \mu, \cdots, x_n - \mu], \tag{7.19}$$

and construct the SVD of $\bar{X}$,

$$\bar{X} = USV^T. \tag{7.20}$$

(a) Show that $\bar{X}$ can be written as

$$\bar{X} = X(I - \frac{1}{n} 1 1^T), \tag{7.21}$$

where 1 is the vector of ones.

(b) Show that $\{u_i, \frac{s_i^2}{n}\}$ is an eigenvector/eigenvalue pair of Σ, where u_i is the ith column of U.

(c) Rewrite the PCA training algorithm to use the SVD.

.....

Problem 7.5 PCA and classification

Consider a classification problem with two classes $y \in \{1, 2\}$ with class probabilities,

$$p(y = 1) = p(y = 2) = 1/2, \tag{7.22}$$

and Gaussian class conditionals,

$$p(x|y = j) = \mathcal{N}(x|\mu_j, \Sigma_j). \tag{7.23}$$

(a) The random variable x is not Gaussian, but we can still compute its mean and covariance. Show that

$$\mu_x = \mathbb{E}[x] = \frac{1}{2}(\mu_1 + \mu_2), \tag{7.24}$$

and

$$\Sigma_x = \mathbb{E}[(x - \mu_x)(x - \mu_x)^T] = \frac{1}{2}(\Sigma_1 + \Sigma_2) + \frac{1}{4}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \tag{7.25}$$

(b) Consider the 2-dimensional case, where

$$\mu_1 = -\mu_2 = \mu = \begin{bmatrix} \alpha \\ 0 \end{bmatrix}, \tag{7.26}$$

with $\alpha > 0$, and

$$\Sigma_1 = \Sigma_2 = \Gamma = \begin{bmatrix} 1 & 0 \\ 0 & \sigma^2 \end{bmatrix}. \tag{7.27}$$

Use the principal component analysis of x to determine the best 1-dimensional subspace, i.e. the transformation

$$z = \phi^T x \tag{7.28}$$

where ϕ is the eigenvector corresponding to the largest eigenvalue of Σ_x. For this problem, is the PCA approach to dimensionality reduction a good one from the classification point of view? Explain why.

.....

Linear Discriminant Analysis

Problem 7.6 Fisher's linear discriminant

Suppose we have feature vectors from two classes, where the sample mean and scatter matrix is given by $\{\mu_1, S_1\}$ for class 1, and $\{\mu_2, S_2\}$ for class 2. The Fisher's linear discriminant (FLD) finds the optimal projection that maximizes the ratio of the "between-class" scatter and the "within-class" scatter,

$$w^* = \underset{w}{\operatorname{argmax}}\, J(w), \quad J(w) = \frac{w^T S_B w}{w^T S_W w}, \tag{7.29}$$

where S_B and S_W are the between- and within-class scatter matrices,

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T, \quad S_W = S_1 + S_2. \tag{7.30}$$

In this problem, we will derive the optimal w^*.

(a) Because (7.29) is maximizing a ratio in $J(w)$, it suffices to fix the denominator to some value and maximize the numerator. Use Lagrange multipliers to rewrite (7.29) as a constrained optimization problem where $w^T S_W w = 1$.

(b) Show that the stationary point of the constrained optimization problem is given by the *generalized* eigenvalue problem,

$$S_B w = \lambda S_W w, \tag{7.31}$$

where λ is the Lagrange multiplier.

(c) Finally, assuming that S_W is invertible, show that the optimal w^* is given by

$$w^* \propto S_W^{-1}(\mu_1 - \mu_2). \tag{7.32}$$

Hint: we are only interested in the direction of the best projection w to maximize $J(w)$; the length of w does not affect the ratio.

.....

Problem 7.7 Fisher's linear discriminant as least-squares regression

In this problem, we will show that Fisher's linear discriminant is equivalent to least-squares regression for a particular set of target values. Let $X = \{x_1, \cdots, x_n\}$ be the data points with $x_i \in \mathbb{R}^d$, and $\mathcal{C}_1$ be the set of points in class 1 (with size n_1), and likewise for class 2 ($\mathcal{C}_2$, n_2). Define the following sample statistics of the data for $j \in \{1, 2\}$,

$$\mu_j = \frac{1}{n_j} \sum_{x_i \in \mathcal{C}_j} x_i, \quad S_j = \sum_{x_i \in \mathcal{C}_j} (x_i - \mu_j)(x_i - \mu_j)^T. \tag{7.33}$$

We want to learn a linear projection of the data to separate the two classes,

$$z_i = w^T x_i + b, \tag{7.34}$$

where $z_i \in \mathbb{R}$ is the projected data point, $w \in \mathbb{R}^d$ is the projection vector and $b \in \mathbb{R}$ is the bias term. We will formulate this as a regression problem. For each data point x_i, we define the target value as the reciprocal of the class probability,

$$t_i = \begin{cases} \frac{n}{n_1}, & y_i = 1 \\ \frac{-n}{n_2}, & y_i = 2. \end{cases} \tag{7.35}$$

The best projection is then obtained by minimizing the least-squares error between the projection z_i and the target t_i,

$$\{w^*, b^*\} = \operatorname*{argmin}_{w,b} E(w, b), \tag{7.36}$$

$$E(w, b) = \frac{1}{2} \sum_{i=1}^{n} (z_i - t_i)^2 = \frac{1}{2} \sum_{i=1}^{n} (w^T x_i + b - t_i)^2. \tag{7.37}$$

This is a linear regression problem where the target values t_i are based on the class of each point.

(a) Show that $\sum_{i=1}^{n} t_i = 0$.

(b) Show that the optimal b^* can be expressed as a function of w,

$$b = -w^T \mu, \qquad \mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(n_1\mu_1 + n_2\mu_2). \tag{7.38}$$

(c) Using (b), show that a stationary point for w satistfies

$$(S_W + \frac{n_1 n_2}{n} S_B)w = n(\mu_1 - \mu_2), \tag{7.39}$$

where S_W and S_B are the within- and between-class matrices,

$$S_W = S_1 + S_2, \qquad S_B = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T. \tag{7.40}$$

(d) Finally, show that the direction of the optimal w can be written as,

$$w^* \propto S_W^{-1}(\mu_2 - \mu_1) \tag{7.41}$$

Hint: note that $S_B w$ is always in the direction of $\mu_2 - \mu_1$.

.....

Probabilistic PCA and Factor Analysis

Problem 7.8 Probabilistic PCA: the model

In this problem, we will consider a more probabilistic version of PCA, called probabilistic PCA (PPCA). Let $x \in \mathbb{R}^d$ be the observation and $z \in \mathbb{R}^k$ the lower-dimensional representation (called a latent variable). We assume the observation x is generated from latent variable z using a linear transformation and mean offset, with some Gaussian observation noise,

$$x = \Phi z + \mu + \epsilon, \tag{7.42}$$

where $\Phi \in \mathbb{R}^{d\times k}$ is the linear transformation, $\mu \in \mathbb{R}^d$ the mean vector, and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ the observation noise. (Note that we do not impose an orthonormal constraint on the columns of Φ). Hence, the conditional distribution of the observation x given z is

$$p(x|z) = \mathcal{N}(x|\Phi z + \mu, \sigma^2 I). \tag{7.43}$$

To complete the probabilistic model, we assume a Gaussian prior distribution on the latent variables,

$$p(z) = \mathcal{N}(z|0, I). \tag{7.44}$$

Note that this probabilistic PCA is a slightly different formulation than standard PCA, in that we are directly modeling the generation of the data from the latent space representation. In contrast, PCA learns the mapping from observations to latent space.

In this problem we will derive the marginal, joint likelihood, and posterior of the PPCA model. In the next problem, we will consider learning the model using maximum likelihood.

(a) Show that the marginal distribution of x is

$$p(x) = \int p(x|z)p(z)dz = \mathcal{N}(x|\mu, \Phi\Phi^T + \sigma^2 I). \tag{7.45}$$

Hint: use Problem 1.9.

(b) Looking at $p(x)$, what is the major difference between the models of PPCA and PCA?

(c) Show that the joint likelihood of x and z is

$$p(x, z) = \mathcal{N}(\begin{bmatrix} x \\ z \end{bmatrix} | \begin{bmatrix} \mu \\ 0 \end{bmatrix}, \begin{bmatrix} \Phi\Phi^T + \sigma^2 I & \Phi^T \\ \Phi & I \end{bmatrix}) \tag{7.46}$$

Hint: define $a = [x^T, z^T]^T$ and find its mean and covariance.

(d) Show that the posterior likelihood of z given x is

$$p(z|x) = \mathcal{N}(z|\mu_{z|x}, \Sigma_{z|x}), \tag{7.47}$$
$$\mu_{z|x} = \Phi^T(\Phi\Phi^T + \sigma^2 I)^{-1}(x - \mu), \tag{7.48}$$
$$\Sigma_{z|x} = I - \Phi^T(\Phi\Phi^T + \sigma^2 I)^{-1}\Phi. \tag{7.49}$$

Hint: complete the square.

(e) Use the matrix inverse identities in Problem 1.15 to show that

$$\mu_{z|x} = M^{-1}\Phi^T(x - \mu), \qquad \Sigma_{z|x} = \sigma^2 M^{-1}, \qquad M = \Phi^T\Phi + \sigma^2 I. \tag{7.50}$$

Hence, calculation of $p(z|x)$ requires inverting only a $k \times k$ matrix, rather than a $d \times d$ matrix.

.....

Problem 7.9 Probabilistic PCA: EM learning

Given a set of data points $X = \{x_1, \cdots, x_n\}$, the goal is to learn the parameters $\{\Phi, \mu, \sigma^2\}$ of the probabilistic PCA model (see Problem 7.8) that maximize the log-likelihood of the data,

$$\log p(X) = \sum_{i=1}^{n} \log p(x_i) = \sum_{i=1}^{n} \log \mathcal{N}(x_i|\mu, \Phi\Phi^T + \sigma^2 I). \tag{7.51}$$

In this case taking the derivative and solving for the parameters will be quite complex (it is possible though). Alternatively, because our model has latent variables $Z = \{z_1, \cdots, z_n\}$ that are unobserved, we can also find the maximum likelihood parameters using the EM algorithm.

(a) Show that the Q function can be written as

$$Q(\theta, \hat{\theta}) = \mathbb{E}_{Z|X,\hat{\theta}}[\log p(X, Z|\theta)] \tag{7.52}$$
$$= -\sum_{i=1}^{n} \left\{ \frac{d}{2}\log\sigma^2 + \frac{1}{2}\mathrm{tr}(\hat{P}_i) + \frac{1}{2\sigma^2}\|x_i - \mu\|^2 - \frac{1}{\sigma^2}\hat{z}_i^T\Phi^T(x_i - \mu) + \frac{1}{2\sigma^2}\mathrm{tr}(\hat{P}_i\Phi^T\Phi) \right\}, \tag{7.53}$$

where the required expectations for the E-step are

$$\hat{z}_i = \mathbb{E}_{Z|X,\hat{\theta}}[z_i] = M^{-1}\Phi^T(x_i - \mu) \tag{7.54}$$

$$\hat{P}_i = \mathbb{E}_{Z|X,\hat{\theta}}[z_i z_i^T] = \sigma^2 M^{-1} + \hat{z}_i \hat{z}_i^T, \tag{7.55}$$

and the parameters used to calculate these expectations are from the old model $\hat{\theta}$.

(b) Show that the M-step is given by

$$\mu_* = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \Phi_* = \left[\sum_{i=1}^{n}(x_i - \mu)\hat{z}_i^T\right]\left[\sum_{i=1}^{n}\hat{P}_i\right]^{-1}, \tag{7.56}$$

$$\sigma_*^2 = \frac{1}{nd}\sum_{i=1}^{n}\left\{\|x_i - \mu\|^2 - 2\hat{z}_i^T\Phi_*^T(x_i - \mu) + \mathrm{tr}(\hat{P}_i\Phi_*^T\Phi_*)\right\}. \tag{7.57}$$

(c) Taking $\sigma^2 \to 0$ yields the standard PCA model. Let $\hat{Z} = [\hat{z}_1, \cdots, \hat{z}_n]$ be the matrix of estimated latent variables, and $\bar{X} = [x_1 - \mu, \cdots, x_n - \mu]$ the matrix of mean-subtracted data points. Show that the resulting EM algorithm for $\sigma \to 0$ is,

$$\text{PCA E} - \text{step}: \;\; \hat{Z} = (\Phi^T\Phi)^{-1}\Phi^T\bar{X} \tag{7.58}$$

$$\text{PCA M} - \text{step}: \;\; \Phi = \bar{X}\hat{Z}^T(\hat{Z}\hat{Z}^T)^{-1} \tag{7.59}$$

Intuitively, what are the E- and M-steps doing in this algorithm?

(d) Analyze the complexity of the EM algorithm for PCA versus using eigen-decomposition for PCA. Recall that computing: k eigenvalues of a $d \times d$ matrix is $O(kd^2)$; the inverse of a $k \times k$ matrix is $O(k^3)$; and the outer product of two vectors ($\mathbb{R}^a$ and $\mathbb{R}^b$) is $O(ab)$. When will it be advantageous to use EM over the standard formulation?

.....

Problem 7.10 Factor Analysis

Factor analysis (FA) is closely related to probabilistic PCA (Problem 7.8). The main difference is that the observation noise is assumed to have a diagonal covariance matrix, rather than an isotropic form, i.e $\epsilon \sim \mathcal{N}(0|\Psi,)$, where Ψ is a diagonal covariance matrix. The conditional distribution of x is then

$$p(x|z) = \mathcal{N}(x|\Phi z + \mu, \Psi). \tag{7.60}$$

In essence, FA is also capturing correlations in the data using Φ, while modeling the observation noise of individual features as independent, but with possibly different variance. In FA literature, the columns of Φ are called *factor loadings*, and the diagonal elements of Ψ are called *uniquenesses*.

Given a set of samples $X = \{x_1, \cdots, x_n\}$, the parameters of the FA model can be learned using the EM algorithm, similar to probabilistic PCA.

(a) Show that the E-step for FA is to calculate:

$$G = (I + \Phi^T\Psi^{-1}\Phi)^{-1}, \tag{7.61}$$

$$\hat{z}_i = G\Phi^T\Psi^{-1}(x_i - \mu), \tag{7.62}$$

$$\hat{P}_i = G + \hat{z}_i\hat{z}_i^T. \tag{7.63}$$

(b) Show that the M-step for FA is:

$$\mu_* = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \Phi_* = \left[\sum_{i=1}^{n}(x_i-\mu)\hat{z}_i^T\right]\left[\sum_{i=1}^{n}\hat{P}_i\right]^{-1}, \tag{7.64}$$

$$\Psi_* = \text{diag}\left\{\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T - \Phi_*\frac{1}{n}\sum_{i=1}^{n}\hat{z}_i(x_i-\mu)^T\right\}, \tag{7.65}$$

where the diag operator sets all non-diagonal elements of a matrix to zero.

.....

Canonical Correlation Analysis

Problem 7.11 Canonical correlation analysis (CCA)

Canonical correlation analysis (CCA) is a dimensionality-reduction method similar to PCA. While PCA deals with only one data space, CCA is method for jointly reducing the dimension across two spaces. Each dimension of the lower-dimensional representation corresponds to *correlated* directions in the data spaces (i.e, the two directions describe the same thing).

Formally, let $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^D$ be a pair of observation random variables (with different dimensions), with covariances matrices

$$\Sigma_{xx} = \mathbb{E}[(x-\mu_x)(x-\mu_x)^T] \tag{7.66}$$

$$\Sigma_{yy} = \mathbb{E}[(y-\mu_y)(y-\mu_y)^T] \tag{7.67}$$

$$\Sigma_{xy} = \mathbb{E}[(x-\mu_x)(y-\mu_y)^T] = \Sigma_{yx}^T, \tag{7.68}$$

where Σ_{xy} is the covariance between the two variables. The covariance matrices can be computed using sample data pairs $\{x_i, y_i\}_{i=1}^n$. Let z_x and z_y be linear projections of x and y,

$$z_x = w_x^T x, \qquad z_y = w_y^T y, \tag{7.69}$$

where $w_x \in \mathbb{R}^d$ and $w_y \in \mathbb{R}^D$ are the projection directions. The goal is to find $\{w_x, w_y\}$ that maximizes the correlation between z_x and z_y,

$$\{w_x^*, w_y^*\} = \underset{w_x\neq 0, w_y\neq 0}{\text{argmax}}\ \rho, \tag{7.70}$$

$$\rho = \frac{\text{cov}(z_x, z_y)}{\sqrt{\text{var}(z_x)\text{var}(z_y)}}. \tag{7.71}$$

In this problem, we will derive the solution using the SVD.

(a) Show that the covariances and variances in (7.71) are:

$$\text{cov}(z_x, z_y) = w_x^T \Sigma_{xy} w_y, \tag{7.72}$$

$$\text{var}(z_x) = w_x^T \Sigma_{xx} w_x, \tag{7.73}$$

$$\text{var}(z_y) = w_y^T \Sigma_{yy} w_y. \tag{7.74}$$

Hence, the correlation function is

$$\rho = \frac{w_x^T \Sigma_{xy} w_y}{\sqrt{w_x^T \Sigma_{xx} w_x}\sqrt{w_y^T \Sigma_{yy} w_y}}. \tag{7.75}$$

(b) Perform a change of basis on (7.75),

$$\hat{w}_x = \Sigma_{xx}^{1/2} w_x, \qquad \hat{w}_y = \Sigma_{yy}^{1/2} w_y, \tag{7.76}$$

where $A^{1/2}$ is the matrix square-root ($A = A^{1/2}A^{1/2}$), and derive an equivalent optimization problem to (7.70),

$$\{u^*, v^*\} = \underset{u,v}{\operatorname{argmax}}\, u^T \Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2} v, \quad \text{s.t. } u^T u = 1,\ v^T v = 1. \tag{7.77}$$

$$w_x^* = \Sigma_{xx}^{-1/2} u^*, \quad w_y^* = \Sigma_{yy}^{-1/2} v^*. \tag{7.78}$$

(c) Consider the SVD decomposition

$$\Sigma_{xx}^{-1/2} \Sigma_{xy} \Sigma_{yy}^{-1/2} = USV^T. \tag{7.79}$$

Show that (7.77) is maximized with $u^* = u_1$ and $v^* = v_1$, where u_1 and v_1 are the left and right singular vectors corresponding to the largest singular value (i.e., the first columns of U and V).

.....