CMSC 35400 / STAT 37710

Spring 2020
Homework 1

Reading assignment: Bishop chapters 1, 2, & 3.

1. Which of the following matrices are positive semi-definite and hence valid covariance matrices?

a) $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

b) $\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$

c) $\begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}$

d) $\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$

SOLUTION: (a), (d) are valid covariance matrices. (b) is not PSD, (c) is not symmetric.

2. Let $D = \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$ where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ are the training data that you are given. As you have to predict a continuous variable, one of the simplest possible model is linear regression.

Consider the following loss function

$$\arg\min_\theta \hat{L}(\theta) = \arg\min_\theta \sum_{i=1}^{n} (y_i - \theta^\top x_i)^2. \tag{1}$$

Let us introduce the $n \times p$ matrix $X \in \mathbb{R}^{n \times p}$ with the x_i as rows and the vector $y \in \mathbb{R}^n$ consisting of the scalars y_i. Then Eq. (1) can be equivalently re-written as

$$\arg\min_\theta \|X\theta - y\|^2$$

We refer to any θ^* that attains the above minimum as a solution to the problem.

a) Show that if $X^\top X$ is invertible, then there is a unique θ^* that can be computed as $\theta^* = (X^\top X)^{-1} X^\top y$.
SOLUTION: Note that

$$\hat{L}(\theta) = \|X\theta - y\|^2 = (X\theta - y)^\top (X\theta - y) = \theta^\top X^\top X\theta - 2\theta^\top X^\top y + y^\top y.$$

The gradient of this function is equal to

$$\nabla \hat{L}(\theta) = 2X^\top X\theta - 2X^\top y.$$

Because $\hat{L}(\theta)$ is convex, its optima are exactly those points that have a zero gradient, i.e., those θ^* that satisfy $X^\top X\theta^* = X^\top y$. Under the given assumption, the unique minimizer is indeed equal to $\theta^* = (X^\top X)^{-1}X^\top y$.

b) Show that for $n < p$, Eq. (1) does not admit a unique solution. Furthermore, intuitively explain why this is the case.
SOLUTION: Consider the SVD $X = U\Sigma V^\top$ where U is an unitary $n \times n$ matrix, V is a unitary $p \times p$ matirx and Σ is a diagonal $n \times p$ matrix, with the singular values of X on the diagonal. We then have

$$\arg\min_\theta \hat{L}(\theta) = \arg\min_\theta \left[\theta^\top V\Sigma^\top \Sigma V^\top \theta - 2y^\top U\Sigma V^\top \theta\right].$$

Rotating θ using V to $z = V^\top \theta$, we get

$$\arg\min_z \left[z^\top \Sigma^\top \Sigma z - 2y^\top U\Sigma z\right] = \arg\min_z \sum_{i=1}^{p} \left[z_i^2\sigma_i^2 - 2(U^\top y)_i z_i \sigma_i\right]$$

where σ_i is the ith entry in the diagonal of Σ. This problem decomposes into p independent optimization problems of the form

$$z_i = \arg\min_z \left[z^2\sigma_i^2 - 2(U^\top y)_i z\sigma_i\right]$$

for $i = 1, \ldots, p$. Therefore, if $\sigma_i \neq 0$, we get

$$z_i = \frac{(U^\top y)_i}{\sigma_i}.$$

For the case $n < p$, X has at most rank n, and hence at most n of its singular values are nonzero. This means that there is at least one index j such that $\sigma_j = 0$ and hence any $z_i \in \mathbb{R}$ is a solution to the optimization problem. As a result, the set of optimal solutions for z, and consequently for w, is a linear subspace of at least one dimension. Therefore, no unique solution exists.
The intuition behind these results is that the linear system $X\theta = y$ is under-determined as there are less data points than parameters that we want to estimate.

3. We observe $z_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$ for $i = 1, \ldots, n$, and consider the problem of estimating μ. We consider some estimators:

a) $\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} z_i$ for some $m < n$. What is the bias of this estimator? What is the variance of this estimator?
SOLUTION: bias $= 0$, variance $= \sigma^2/m$

b) $\hat{\mu}_0 = 0$. What is the bias of this estimator? What is the variance of this estimator?
SOLUTION: bias $= \mu$, variance $= 0$.

c) $\hat{\mu}_{\lambda,m} = \lambda\hat{\mu}_m + (1-\lambda)\hat{\mu}_0$ for some $\lambda \in (0,1)$. What is the bias of this estimator? What is the variance of this estimator? What is the best (in terms of minimizing MSE) value of λ for given values of μ and m?
SOLUTION: squared bias $= (1-\lambda)^2\mu^2$. variance $= \lambda^2\sigma^2/m$. MSE $= (1-\lambda)^2\mu^2 + \lambda^2\sigma^2/m$. Taking the derivative with respect to λ and setting it to zero, we find $\lambda^* = \frac{m\mu^2}{\sigma^2+1}$. So when m is larger, we should favor $\hat{\mu}_m$ more, but when μ is smaller, we should favor $\hat{\mu}_0$ more.

4. Consider a random observation vector

$$y = X\theta + \epsilon.$$

where X is an $n \times p$, $p \ll n$, deterministic matrix and

$$\theta \sim \mathcal{N}(0, \sigma_\theta^2 I_p).$$

ϵ is a random noise vector, independent of θ given by

$$\epsilon \sim N(0, \sigma_\epsilon^2 I_n).$$

a) Give an expression for the covariance R_{yy} of y in terms of X.
SOLUTION: First note that

$$\mathbb{E}y = \mathbb{E}X\theta + \mathbb{E}\epsilon = 0.$$

Then

$$\begin{aligned}
R_{yy} &= \mathbb{E}[(y-\mathbb{E}[y])(y-\mathbb{E}[y])^\top] \\
&= \mathbb{E}[(X\theta+\epsilon)(X\theta+\epsilon)^\top] \\
&= \mathbb{E}[X\theta(X\theta)^\top] + \mathbb{E}[X\theta\epsilon^\top] + \mathbb{E}[\epsilon(X\theta)^\top] + \mathbb{E}[\epsilon\epsilon^\top] \\
&= X\mathbb{E}[\theta\theta^\top]X^\top + \mathbb{E}[X\theta\epsilon^\top] + \mathbb{E}[\epsilon(X\theta)^\top] + \mathbb{E}[\epsilon\epsilon^\top] \\
&= \sigma_\theta^2 XX^\top + \sigma_\epsilon^2 I_n + \mathbb{E}[X\theta]\mathbb{E}[\epsilon^\top] + \mathbb{E}[\epsilon]\mathbb{E}[x^\top] \\
&= \sigma_\theta^2 XX^\top + \sigma_\epsilon^2 I_n.
\end{aligned}$$

b) Assuming the p columns of X are orthonormal vectors, determine the first p eigenvalues and eigenvectors of R_{yy}. How are the eigenvectors and eigenvalues related to X? (HINT: What can we do to R_{yy} that exploits the fact that the columns of X are orthonormal?)
SOLUTION: First recall that an eigenvector v of a matrix A satisfies $Av = \lambda v$, where λ, a scalar, is the eigenvalue of A associated with the eigenvector v. Then we examine the following:

$$R_{yy}X = \sigma_\theta^2 XX^\top X + \sigma_\epsilon^2 I_n X = \sigma_\theta^2 X + \sigma_\epsilon^2 X = (\sigma_\theta^2 + \sigma_\epsilon^2)X,$$

where we use the fact that the columns of X are orthonormal. So we see that the columns of X are first p eigenvectors of R_{yy} with eigenvalues $(\sigma_\theta^2 + \sigma_\epsilon^2)$.

c) Let's put the ideas above into action. Generate 1000 random signals in noise and form the *sample covariance matrix* S according to the MATLAB code below:

```
M = 10000;
n = 32;
p = 2;
sig_t = .1;
sig_e = .01;
X = ones(n,p);
for f = 0:(p-2)
  X(:,f+2) = kron(ones(2^f,1),kron([1 -1]',ones(n/2^(f+1),1)));
end
X = normc(X);
S = zeros(n);
for m = 1:M
  theta = randn(p,1)*sig_t;
  y = X*theta + randn(n,1)*sig_e;
  S = S + y*y';
end;
S = S/M;
```

Use the built-in eigenvalue and eigenvector function (`eig` in MATLAB) to determine a small set of vectors that span a subspace that contains most of the data. How are these related to the columns of X?

SOLUTION:

```
[V,Lambda] = eigs(S,2);
figure(3);clf; plot(V);
```

The columns of V span the same subspace as the columns of X.