Positive Semidefinite (PSD) Matrices

Principal Component Analysis (PCA)

Minimum Mean Squared Error (MMSE) Estimation

ECE 285: Weeks 9 – 10

Piya Pal

Department of Electrical and Computer Engineering
University of California, San Diego

Winter 2017
Outline

1. Positive Semidefinite (PSD) Matrices
2. Principal Component Analysis (PCA)
3. Minimum Mean Squared Error (MMSE) Estimation
Positive Semidefinite Matrices

Definition

A Hermitian matrix $A \in \mathbb{C}^{n \times n}$ is positive semidefinite (PSD) if

$$x^H Ax \geq 0, \quad \forall x \in \mathbb{C}^n$$

It is positive definite (PD) if

$$x^H Ax > 0, \quad \forall x \in \mathbb{C}^n, x \neq 0$$

- A Hermitian $A \in \mathbb{C}^{n \times n}$ is PSD if and only if its eigenvalues are non-negative. Similarly it is PD if and only if its eigenvalues are positive. Hence, a PD matrix is nonsingular.
- A PSD $A$ can always be factored as $A = B^H B$. The factor $B$ is non unique. If $A = U \Lambda U^H$ represents the eigenvalue decomposition of $A$ (where $U$ is unitary and $\Lambda$ is diagonal with non negative entries), then one choice of $B = U \Lambda^{1/2}$. 
PSD matrices are associated with the idea of a “generalized inequality”. Recall the standard inequality for real-valued scalars $x, y \in \mathbb{R}$, whereby we can compare them as $x \geq y$ (or vice versa). For two PSD matrices $A, B \in \mathbb{C}^n$, a generalized inequality $A \succeq B$ is defined as follows:

$$A \succeq B \iff A - B \succeq 0 \iff x^H (A - B) x \geq 0, \forall x \in \mathbb{C}^n$$

Notice that for a Hermitian $A$, the scalar function $f(x) = x^H A x$ is real-valued and a quadratic function of $x$. When $A$ is PSD, $f(x)$ is non-negative, i.e. $f(x) \geq 0, \forall x$. 
Theorem

Consider a Hermitian matrix $P$ partitioned as

$$P = \begin{pmatrix} A & B \\ B^H & C \end{pmatrix}$$

Then

$$P \succ 0 \iff A \succ 0 \text{ and } C - B^H A^{-1} B \succ 0$$

The matrix $C - B^H A^{-1} B$ is called the Schur Complement of $A$ in $P$. The Schur Complement of $C$ can be similarly defined.
Covariance Matrix

Given a random vector $\mathbf{x} \in \mathbb{R}^n$ with probability density function $f_x(\mathbf{x})$ (which is the joint pdf of the $n$ elements of $\mathbf{x}$), its mean ($\mu_x$) and covariance matrix ($\Sigma_{xx}$) are defined as

$$
\mu_x = E(\mathbf{x}) = \int_{\mathbb{R}^m} x f_x(\mathbf{x}) d\mathbf{x}
$$

$$
\Sigma_{xx} = E(\mathbf{x} - \mu_x)(\mathbf{x} - \mu_x)^T = \int_{\mathbb{R}^m} (\mathbf{x} - \mu_x)(\mathbf{x} - \mu_x)^T f_x(\mathbf{x}) d\mathbf{x}
$$

The covariance matrix $\Sigma_{xx}$ is a PSD matrix. Given random vectors $\mathbf{x} \in \mathbb{R}^m, \mathbf{y} \in \mathbb{R}^n$ with respective means $\mu_x$ and $\mu_y$, their cross covariance matrix $\Sigma_{xy} \in \mathbb{R}^{m \times n}$ is defined as

$$
\Sigma_{xy} = E(\mathbf{x} - \mu_x)(\mathbf{y} - \mu_y)^T
$$

The expectation is taken over the joint pdf $f_{x,y}(\mathbf{x}, \mathbf{y})$ of $\mathbf{x}$ and $\mathbf{y}$.
Consider $n$ zero-mean random variables $y_1, y_2, \cdots, y_n \in \mathbb{R}$. They are said to be uncorrelated if $E(y_i y_j) = 0$, $\forall i, j$. In this case the covariance matrix of $y = [y_1, y_2, \cdots, y_n]^T$ is a **diagonal matrix**.

Given $n$ zero-mean correlated random variables $x_1, x_2, \cdots, x_n$, the covariance matrix of $x = [x_1, \cdots, x_n]^T$, denoted by $\Sigma_{xx}$, is non-diagonal. Since $\Sigma_{xx}$ is PSD, it can be diagonalized as $\Sigma_{xx} = U \Lambda U^H$ where $U$ is unitary and $\Lambda$ is diagonal with non-negative elements. Consider the random vector

$$y = U^H x$$

It can be easily verified that the covariance matrix of $y$ is the diagonal matrix $\Lambda$. Hence the elements of $y$ are uncorrelated. This process is called “whitening”, and forms the key idea behind Principal Component Analysis (PCA).
**Principal Component Analysis (PCA)**

PCA is a data-driven technique, where an idea similar to whitening is applied, using the *sample covariance matrix* as an estimate of the true covariance matrix.

- Consider $n$ data vectors $x_i \in \mathbb{R}^m$, $i = 1, 2, \cdots, n$, whose sample mean $\hat{\mu}_x = \frac{1}{n} \sum_{i=1}^n x_i$ is zero. Let $X = [x_1, x_2, \cdots, x_n]$ be the $m \times n$ data matrix. \(^1\)

- The unbiased sample covariance matrix (SCM) of the data is given by

$$\hat{\Sigma}_{xx} = \frac{1}{n-1} \sum x_i x_i^T = \frac{1}{n-1} XX^T$$

Notice that $\hat{\Sigma}_{xx}$ is PSD. The key idea behind PCA is that in many cases, the rank $r$ of $\hat{\Sigma}_{xx}$ is much smaller than the dimension $m$. Hence $\hat{\Sigma}_{xx}$ has only $r \ll m$ positive eigenvalues.

\(^1\)Otherwise, we can always calculate $\hat{\mu}_x$ from the data and subtract it from each $x_i$.
PCA (contd.)

- The SCM $\hat{\Sigma}_{xx}$ can be diagonalized as $\hat{\Sigma}_{xx} = U \Lambda U^H$ where $U$ is unitary and the diagonal matrix $\Lambda$ has exactly $r$ non-zero elements.
- Consider the transformation
  \[ y_i = U_r^H x_i \]
  where $U_r \in \mathbb{R}^{m \times r}$ is obtained by partitioning $U$ as $U = [U_r, U_{m-r}]$.
- Verify that the SCM of the transformed data $y_1, y_2, \cdots, y_n$ is a diagonal matrix containing the $r$ non-zero entries of $\Lambda$.
- When $r \ll m$, PCA performs dimension reduction by transforming higher dimensional data $x_i$ into lower dimensional representations $y_i$.  

\[ ^2 \text{You can also retain } k < r \text{ principal components of the data, but in this case, you cannot recover the data covariance matrix} \]
Suppose we want to estimate a random vector $\mathbf{x} \in \mathbb{R}^m$ given the random vector $\mathbf{y} \in \mathbb{R}^n$, i.e., determine a suitable function $h(\mathbf{y}) \in \mathbb{R}^m$ of $\mathbf{y}$. Note that $h(\mathbf{y})$ need not be linear. The MMSE estimator of $\mathbf{x}$, given the observation $\mathbf{y}$, is defined as

$$h_{\text{MMSE}}(\mathbf{y}) = \arg \min_{h(.)} \mathbb{E} \left[ (\mathbf{x} - h(\mathbf{y}))^T (\mathbf{x} - h(\mathbf{y})) \right]$$  \hspace{1cm} (1)$$

Notice that in (1), the expectation is over the joint pdf of $\mathbf{x}$ and $\mathbf{y}$, i.e. $h_{\text{MMSE}}(\mathbf{y})$ is that function which minimizes the squared error averaged over all possible realizations of the random variables $(\mathbf{x}, \mathbf{y})$. 
Theorem

An estimator $h(y)$ of $x$ given $y$ is an MMSE estimator if and only if the associated error $x - h(y)$ is orthogonal to any (vector-valued) function $g(y)$ of $y$, i.e.

$$E[(x - h(y))g^T(y)] = 0$$

It can be shown that the MMSE estimator of $x$ given $y$ is

$$h_{MMSE}(y) = E(x|y)$$

where $E(x|y) = \int_{\mathbb{R}^m} xf_{x|y}(x|y)dx$ and $f_{x|y}(x|y) = \frac{f_{x,y}(x,y)}{f(y)}$
A random vector $\mathbf{x} \in \mathbb{R}^m$ with mean $\mu_x$ and covariance matrix $\Sigma_{xx} \succ 0$ is said to follow a multivariate Gaussian distribution, i.e. $\mathbf{x} \sim \mathcal{N}(\mu_x, \Sigma_{xx})$ if

$$f_x(\mathbf{x}) = \frac{1}{(2\pi)^{m/2} \sqrt{\det \Sigma_{xx}}} e^{-\frac{1}{2} (\mathbf{x} - \mu_x)^T \Sigma_{xx}^{-1} (\mathbf{x} - \mu_x)}$$

- If $\mathbf{x} = [x_1, x_2]^T$, then
  $$x_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$$

  where $\mu_x = [\mu_1, \mu_2]^T$ and $\Sigma_{xx} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$

- A similar result holds for $x_2$. 
Multivariate Gaussian and MMSE Estimation

- If \( x \in \mathbb{R}^m \) and \( y \in \mathbb{R}^n \) are jointly Gaussian with mean
  \[ \mu = [\mu_x, \mu_y]^T \]
  and covariance matrix \( \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix} \)
  i.e.
  \[ f_{x,y}(x, y) = \mathcal{N}(\mu, \Sigma) \]
  then
  \[ f_{x|y}(x|y) = \mathcal{N}(\mu_{x|y}, \Sigma_{x|y}) \]

  \[ \mu_{x|y} = \mu_x + \Sigma_{xy} \Sigma_{yy}^{-1} (y - \mu_y), \quad \Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} \]

- The MMSE estimator of \( x \) given \( y \) is therefore given by
  \[ h_{MMSE}(y) = E(x|y) = \mu_{x|y} = \mu_x + \Sigma_{xy} \Sigma_{yy}^{-1} (y - \mu_y) \]
  which is an affine (linear, if \( x, y \) are zero-mean) function of \( y \).