3.3 Variational Characterization of Singular Values

Since the singular values are square roots of the eigenvalues of the Hermitian matrices $A^*A$ and $AA^*$, the singular values inherit the variational characterizations that were explored in Section 2.2. For example,

$$
\sigma_1 = \max_{v \in \mathbb{C}^n} \left( \frac{v^*A^*Av}{v^*v} \right)^{1/2} = \max_{u \in \mathbb{C}^m} \left( \frac{u^*AA^*u}{u^*u} \right)^{1/2},
$$

with the leading right and left singular vectors $v_1$ and $u_1$ being unit vectors that attain these maxima.

However, the singular values also satisfy a subtler variational property that incorporates both left and right singular vectors at the same time. Consider, for unit vectors $u \in \mathbb{C}^m$ and $v \in \mathbb{C}^n$, the quantity

$$
|u^*Av| \leq \|u\| \|A\| \|v\| = \sigma_1,
$$

using the Cauchy–Schwarz inequality and the definition of the induced matrix 2-norm. On the other hand, if $u_1$ and $v_1$ unit vectors that give

$$
|u_1^*A v_1| = |u_1^* (\sigma_1 u_1)| = \sigma_1.
$$

Hence

$$
\sigma_1 = \max_{u \in \mathbb{C}^m, v \in \mathbb{C}^n} \frac{u^*Av}{\|u\| \|v\|}.
$$

We can characterize subsequent singular values the same way. Recall the dyadic version of the singular value decomposition,

$$
A = \sum_{j=1}^{r} \sigma_j u_j v_j^*.
$$

for $A$ of rank $r$. If we restrict unit vectors $u$ and $v$ such to be orthogonal to $u_1$ and $v_1$, then

$$
u^*Av = \sum_{j=1}^{r} \sigma_j u_j^* v_j^* v = \sum_{j=2}^{r} \sigma_j u_j^* v_j^* v = u^* \left( \sum_{j=2}^{r} \sigma_j u_j v_j^* \right) v.
$$

Hence

$$
|u^*Av| \leq \sigma_2,
$$

with the inequality attained when $u = u_1$ and $v = v_1$. Continuing this process gives the following analogue of Theorem 2.2.
Theorem 3.4. For any $A \in \mathbb{C}^{m \times n}$,

$$\sigma_k = \max_{\substack{u \perp \text{span}\{u_1, \ldots, u_{k-1}\} \\ v \perp \text{span}\{v_1, \ldots, v_{k-1}\}}} \frac{|u^*Av|}{\|u\|\|v\|}.$$ 

3.4 Principal Component Analysis

Matrix theory enables the analysis of the volumes of data that now so commonly arise from applications ranging from basic science to public policy. Such measured data often depends on many factors, and we seek to identify those that are most critical. Within this realm of multivariate statistics, principal component analysis (PCA) is a fundamental tool.

Linear algebraists often say, “PCA is the SVD” – in this section, we will explain what this means, and some of the subtleties involved.

3.4.1 Variance and covariance

To understand principal component analysis, we need some basic notions from statistics, described in any basic textbook. For a general description of PCA along with numerous applications, see the text by Jolliffe [Jol02], whose presentation shaped parts of our discussion here.

The expected value, or mean, of a random variable $X$ is denoted $E[X]$. The expected value is a linear function, so for any constants $\alpha, \beta \in \mathbb{R}$, $E[\alpha X + \beta] = \alpha E[X] + \beta$.

The variance of $X$ describes how much $X$ is expected to deviate from its mean,

$$\text{Var}(X) = E[(X - E[X])^2],$$

which, using linearity of the expected value, takes the equivalent form

$$\text{Var}(X) = E[X^2] - E[X]^2.$$ 

The covariance between two (potentially correlated) random variables $X$ and $Y$ is


with $\text{Cov}(X, X) = \text{Var}(X)$. These definitions of variance and covariance are the bedrock concepts underneath PCA, for with them we can understand the variance present in a linear combination of several random variables.
3.4. Principal Component Analysis

Suppose we have a set of real-valued random variables $X_1, \ldots, X_n$ in which we suspect there may be some redundancy. Perhaps some of these variables can be expressed as linear combinations of the others—either exactly, or nearly so. At the other extreme, there may be some way to combine $X_1, \ldots, X_n$ that captures much of the variance in one (or a few) aggregate random variables. In particular, we shall seek scalars $\gamma_1, \ldots, \gamma_n$ such that

$$\sum_{j=1}^n \gamma_j X_j$$

has the largest possible variance. The definitions of variance and covariance, along with the linearity of the expected value, lead to a formula for the variance of a linear combination of random variables:

$$\text{Var}\left(\sum_{j=1}^n \gamma_j X_j\right) = \sum_{j=1}^n \sum_{k=1}^n \gamma_j \gamma_k \text{Cov}(X_j, X_k).$$

You have seen double sums like this before. If we define the covariance matrix $\mathbf{C} \in \mathbb{C}^{n \times n}$ having $(j, k)$ entry

$$c_{j,k} = \text{Cov}(X_j, X_k),$$

and let $\mathbf{v} = [\gamma_1, \ldots, \gamma_n]^T$, then the variance of the combined variable is just a Rayleigh quotient:

$$\text{Var}\left(\sum_{j=1}^n \gamma_j X_j\right) = \mathbf{v}^* \mathbf{C} \mathbf{v}.$$

Since the covariance function is symmetric: $\text{Cov}(X, Y) = \text{Cov}(Y, X)$, the matrix $\mathbf{C}$ is Hermitian; it is also positive semidefinite. Why? Variance, by its definition as the expected value of the square of a real random variable, is always nonnegative. Thus the formula (3.1), which derives from the linearity of the expected value, ensures that $\mathbf{v}^* \mathbf{C} \mathbf{v} \geq 0$. (Under what circumstances can this quantity be zero?)

We can write $\mathbf{C}$ in another convenient way. Collect the random variables into the vector

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}.$$

Then the $(j, k)$ entry of $\mathbb{E}[\mathbf{X}\mathbf{X}^*] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}^*]$ is

$$\mathbb{E}[X_j X_k] - \mathbb{E}[X_j]\mathbb{E}[X_k] = \text{Cov}(X_j, X_k) = c_{j,k},$$

and so

$$\mathbf{C} = \mathbb{E}[\mathbf{X}\mathbf{X}^*] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}^*].$$

Embree – draft – 7 March 2012
3.4.2 Derived variables that maximize variance

Return now to the problem of maximizing the variance of $v^* C v$. Without constraint on $v$, this quantity can be arbitrarily large (assuming $C$ is nonzero); thus we shall require that $\sum_{j=1}^{k} \gamma_j^2 = \|v\|^2 = 1$. With this normalization, you immediately see how to maximize the variance $v^* C v$: $v$ should be a unit eigenvector associated with the largest magnitude eigenvalue of $C$; call this vector $v_1$. The associated variance, of course, is the largest eigenvalue of $C$; call it

$$\lambda_1 = v_1^* C v_1 = \max_{v \in \mathbb{C}^n} \frac{v^* C v}{v^* v}.$$ 

The eigenvector $v_1$ encodes the way to combine $X_1, \ldots, X_n$ to maximize variance. The new variable – the leading principal component – is

$$v_1^* X = \sum_{j=1}^{n} \gamma_j X_j.$$ 

You are already suspecting that a unit eigenvector associated with the second largest eigenvalue, $v_2$ with $\lambda_2 = v_2^* C v_2$, must encode the second-largest way to maximize variance.

Let us explore this intuition. To find the second-best way to combine the variables, we should insist that the next new variable, for now call it $w^* X$, should be independent of the first, i.e.,

$$\text{Cov}(v_1^* X, w^* X) = 0.$$ 

However, using linearity of expectation and the fact that, e.g., $w^* X = X^* w$ for real vectors,

$$\text{Cov}(v_1^* X, w^* X) = E[(v_1^* X)(w^* X)] - E[v_1^* X]E[w^* X]$$
$$= E[(v_1^* X)(w^* X)] - E[v_1^* X]E[X^* w]$$
$$= v_1^* E[XX^*]w - v_1^* E[X]E[X^*]w$$
$$= v_1^* (E[XX^*] - E[X]E[X^*])w$$
$$= v_1^* Cw = \lambda_1 v_1^* w.$$ 

Hence (assuming $\lambda_1 \neq 0$), for the combined variables $v_1^* X$ and $w^* X$ to be independent, the vectors $v_1$ and $w$ must be orthogonal, perfectly confirming your intuition: the second-best way to combine the variables is to pick
3.4. Principal Component Analysis

\textbf{w} to be a unit eigenvector \textbf{v}_2 of \textbf{C} corresponding to the second largest eigenvalue – a direct result of the variational characterization of eigenvalues in Theorem 2.2. The associated variance of \textbf{v}_2^\text{T}\textbf{X} is

\[
\lambda_2 = \max_{\textbf{w} \perp \text{span}\{\textbf{u}_1\}} \frac{\textbf{w}^\text{T}\textbf{C}\textbf{w}}{\textbf{w}^\text{T}\textbf{w}}.
\]

Of course, in general, the \textit{kth} best way to combine the variables is given by the eigenvector \textbf{v}_k of \textbf{C} associated with the \textit{kth} largest eigenvalue.

We learn much about our variables from the relative size of the variances (eigenvalues)

\[
\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0.
\]

If some of the latter eigenvalues are very small, that indicates that the set of \textit{n} random variables can be well approximated by a fewer number of aggregate variables. These aggregate variables are the \textit{principal components} of \textit{X}_1, \ldots, \textit{X}_n.

\subsection*{3.4.3 Approximate PCA from empirical data}

In practical situations, one often seeks to analyze empirical data drawn from some unknown distribution: the expected values and covariances are not available. Instead, we will \textit{estimate} these from the measured data.

Suppose, as before, that we are considering \textit{n} random variables, \textit{X}_1, \ldots, \textit{X}_n, with \textit{m} samples of each:

\[
x_{j,k}, \quad k = 1, \ldots, m,
\]

i.e., \textit{x}_{j,k} is the \textit{kth} sample of the random variable \textit{X}_j. The expected value has a the familiar \textit{unbiased estimate}

\[
\mu_j = \frac{1}{m} \sum_{j=1}^{m} x_{j,k}.
\]

Similarly, we can approximate the covariance

\[
\text{Cov}(X_j, X_k) = \mathbb{E}[(X_j - \mathbb{E}[X_j])(X_k - \mathbb{E}[X_k])].
\]

One might naturally estimate this as

\[
\frac{1}{m} \sum_{\ell=1}^{m} (x_{j,\ell} - \mu_j)(x_{k,\ell} - \mu_k).
\]
However, replacing the true expected values $E[X_j]$ and $E[X_k]$ with the empirical estimates $\mu_j$ and $\mu_k$ introduces some slight bias into this estimate. This bias can be removed by applying Bessel’s correction [??], replacing $1/m$ by $1/(m-1)$ to get the unbiased estimate

$$s_{j,k} = \frac{1}{m-1} \sum_{\ell=1}^{m} (x_{j,\ell} - \mu_j)(x_{k,\ell} - \mu_k), \quad j, k = 1, \ldots, n.$$

If we let

$$\mathbf{x}_j = \begin{bmatrix} x_{j,1} \\ \vdots \\ x_{j,m} \end{bmatrix}, \quad j = 1, \ldots, n,$$

then each covariance estimate is just an inner product

$$s_{j,k} = \frac{1}{m-1}(\mathbf{x}_j - \mu_j)^*(\mathbf{x}_k - \mu_k).$$

Thus, if we center the samples of each variable about its empirical mean, we can write the empirical covariance matrix $\mathbf{S} = [s_{j,k}]$ as a matrix product. Let

$$\mathbf{X} := [(x_1 - \mu_1) \quad (x_2 - \mu_2) \quad \cdots \quad (x_n - \mu_n)] \in \mathbb{R}^{m \times n},$$

so that

$$\mathbf{S} = \frac{1}{m-1} \mathbf{X}^* \mathbf{X}.$$

Now conduct principal component analysis just as before, but with the empirical covariance matrix $\mathbf{S}$ replacing the true covariance matrix $\mathbf{C}$. The eigenvectors of $\mathbf{S}$ now lead to sample principal components. Note that there is no need to explicitly form the matrix $\mathbf{S}$: instead, we can simply perform the singular value decomposition of the data matrix $\mathbf{X}$. This is why some say, “PCA is just the SVD.” We summarize the details step-by-step.

1. Collect $m$ samples of each of $n$ random variables, $x_{j,k}$ for $j = 1, \ldots, m$ and $k = 1, \ldots, n$. (We need $m > 1$, and, generally expect $m \gg n$.)

2. Compute the empirical means, $\mu_k = (\sum_{j=1}^{m} x_{j,k})/m$.

3. Stacking the samples of the $k$th variable in the vector $\mathbf{x}_k \in \mathbb{R}^m$, construct the mean-centered data matrix

$$\mathbf{X} = [(x_1 - \mu_1) \quad (x_2 - \mu_2) \quad \cdots \quad (x_n - \mu_n)] \in \mathbb{R}^{m \times n}.$$
4. Compute the (skiny) singular value decomposition \( \mathbf{X} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^* \), with 
\( \mathbf{U} \in \mathbb{R}^{m \times n} \), \( \mathbf{\Sigma} = \text{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n \times n} \), and \( \mathbf{V} = [\mathbf{v}_1 \cdots \mathbf{v}_n] \in \mathbb{R}^{n \times n} \).

5. The \( k \)th sample principal component is given by \( \mathbf{v}_k^* \mathbf{X} \), where \( \mathbf{X} = [X_1, \ldots, X_n]^* \) is the vector of random variables.

A word of caution: when conducting principal component analysis, the scale of each column matters. For example, if the random variables sampled in each column of \( \mathbf{X} \) are measurements of physical quantities, they can differ considerably in magnitude depending on the units of measurement. By changing units of measurement, you can significantly alter the principal components.