Chapter 5

The Best Basis

5.1 Introduction

The problem we want to consider is this: We're given p points in $\mathbb{R}^n$. Find the "best" $k-$dimensional basis for the data.

There are a couple of things that will make our job easier:

- We will assume that the data has been mean-subtracted, so that the mean is zero (in $\mathbb{R}^n$).
- The basis is orthonormal (each basis vector is in $\mathbb{R}^n$).
- To find the "best" basis will require an error function. We will then minimize it.

At the end of this section, you'll see that the best $k-$dimensional basis for your data (regardless of k) is given by the first k **eigenvectors of the covariance matrix**, which are typically computed using the Singular Value Decomposition (SVD).

5.1.1 The Covariance Matrix, Revisited

Suppose we have p data points in $\mathbb{R}^n$, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p\}$, and they are organized column-wise in an $n \times p$ matrix X.

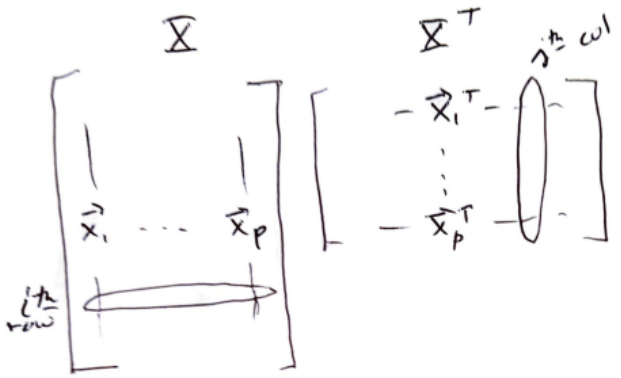
As we recall, the $n \times n$ covariance matrix for data in $\mathbb{R}^n$ measures the covariance between the data in coordinate i and the data in coordinate j. Using the $n \times p$ matrix X, then define $\bar{\boldsymbol{x}} \in \mathbb{R}^n$ as the mean, then the $(i,j)^{\text{th}}$ entry of the covariance matrix is given by the following, where we're taking the covariance between the i^{th} and j^{th} row of X.

$$C_{ij} = \frac{1}{p-1} \sum_{k=1}^{p} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)$$

We will typically assume the mean is zero, so be sure and mean-subtract your data matrix before finding a basis for your data! With zero mean,

$$C_{ij} = \frac{1}{p-1} \sum_{k=1}^{p} x_{ik} x_{jk}$$

If we think of this computation in terms of the matrix X as in the figure below, we see that C_{ij} can be computed using a dot product between row i of X and column j of X^T:



Therefore, we see that the $n \times n$ matrix C can be computed one of two equivalent ways:

$$C = \frac{1}{p-1} XX^T \qquad \text{or} \qquad C = \frac{1}{p-1} \sum_{k=1}^{n} \boldsymbol{x}_k \boldsymbol{x}_k^T$$

Finally, we recognize that the covariance matrix is symmetric, so the **Spectral Theorem** applies. In particular, there is an orthonormal matrix P and a diagonal matrix D so that

$$C = PDP^T$$

where the columns of P form the eigenvectors associated with the diagonal elements of D (which are typically written largest to smallest).

To connect this to the SVD of the data matrix X, if X is $n \times p$ (so that data is stored column-wise), and we write the **reduced** SVD as:

$$X = U\Sigma V^T$$

Then

$$C = \frac{1}{p-1} XX^T = \frac{1}{p-1} U\Sigma V^T V \Sigma^T U^T = U \left(\frac{1}{p-1} \Sigma^2 \right) U^T$$

We see a relationship between the singular values of X, σ_i, and the eigenvalues of the covariance matrix, $\hat{\lambda}_i$:

$$\frac{1}{p-1} \sigma_i^2 = \hat{\lambda}_i$$

So far, we have defined a data matrix X, and we've looked at its covariance matrix C, and we've discovered that the eigenvectors of the covariance matrix are the left singular vectors of the data matrix (when the data is written column-wise and has been mean subtracted).

We'll be getting back to the best basis in a moment, but first we want to make a few more observations.

Projections and the Mean

Suppose you have your p data points in $\mathbb{R}^n$ that have *not* been mean subtracted, and you have a vector $\boldsymbol{u}$ onto which we want to project the data.

First, if we project one point $\boldsymbol{x}$ onto our (unit) vector $\boldsymbol{u}$, then the projection is $(\boldsymbol{x}^T\boldsymbol{u})\boldsymbol{u}$, and the scalar projection is the number $(\boldsymbol{x}^T\boldsymbol{u})$. Similarly, projecting all the data gives us p real numbers (the scalar projections):

$$\left\{ \boldsymbol{u}^T\boldsymbol{x}_1, \boldsymbol{u}^T\boldsymbol{x}_2, \ldots, \boldsymbol{u}^T\boldsymbol{x}_p \right\}$$

so the mean of the projected data is given by

$$\frac{1}{p} \left(\boldsymbol{u}^T\boldsymbol{x}_1 + \boldsymbol{u}^T\boldsymbol{x}_2 + \ldots + \boldsymbol{u}^T\boldsymbol{x}_p \right) = \boldsymbol{u}^T \frac{(\boldsymbol{x}_1 + \boldsymbol{x}_2 + \ldots + \boldsymbol{x}_p)}{p} = \boldsymbol{u}^T \bar{\boldsymbol{x}}$$

Therefore, **the mean of the projection is the projection of the mean**. In particular, if the mean of a data set is zero, and the data is projected to a vector (or subspace), then the new mean is also zero.

Projections and the Variance

In the last section, we saw how the mean and the projection interacted. In this section, let's see how the variance is affected. We'll keep our previous general data set, p points in $\mathbb{R}^n$, and we'll suppose that **the data has zero mean** in $\mathbb{R}^n$. By what we showed, the scalar projections would also have zero mean.

If we project the p vectors onto an arbitrary unit vector $\boldsymbol{u}$, and consider the **scalar projections**, then the resulting variance (in the direction of $\boldsymbol{u}$ will be:

$$S_u^2 = \frac{1}{p-1}\sum_{k=1}^{p}(\boldsymbol{u}^T\boldsymbol{x}_k)^2 = \frac{1}{p-1}\sum_{k=1}^{p}\boldsymbol{u}^T\boldsymbol{x}_k\boldsymbol{x}_k^T\boldsymbol{u} = \boldsymbol{u}^T\left(\frac{1}{p-1}\sum_{k=1}^{p}\boldsymbol{x}_k\boldsymbol{x}_k{}^T\right)\boldsymbol{u} = \boldsymbol{u}^T C\boldsymbol{u}$$

This is actually a very key quantity, and will come up in the next section. We will look at this quantity more closely in a bit, but let's look at what happens in one **special case**: Suppose that $\boldsymbol{u}$ is an eigenvector of the covariance C corresponding to the first eigenvalue, $\hat{\lambda}_1$ using our previous notation. Then:

$$\boldsymbol{u}^T C\boldsymbol{u} = \boldsymbol{u}^T\hat{\lambda}_1\boldsymbol{u} = \hat{\lambda_1}$$

Therefore, if we project all the data to the first eigenvector of C, the new variance will be the first eigenvalue of C.

Reconstruction Error and the Basis

Given a specific vector $\boldsymbol{x} \in \mathbb{R}^n$ and an arbitrary orthonormal basis, $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_n$, we can write

$$\boldsymbol{x} = (\boldsymbol{\phi}_1^T\boldsymbol{x})\boldsymbol{\phi}_1 + (\boldsymbol{\phi}_2^T\boldsymbol{x})\boldsymbol{\phi}_2 + \ldots + (\boldsymbol{\phi}_n^T\boldsymbol{x})\boldsymbol{\phi}_n$$

so that the magnitude of $\boldsymbol{x}$ can be written as:

$$\|\boldsymbol{x}\|^2 = (\boldsymbol{\phi}_1^T\boldsymbol{x})^2 + (\boldsymbol{\phi}_2^T\boldsymbol{x})^2 + \ldots + (\boldsymbol{\phi}_n^T\boldsymbol{x})^2.$$

We can use the same algebraic manipulation that we used in the last section to rewrite this as:

$$\|\boldsymbol{x}\|^2 = \boldsymbol{\phi}_1^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{\phi}_1 + \boldsymbol{\phi}_2^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{\phi}_2 + \ldots + \boldsymbol{\phi}_n^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{\phi}_n.$$

We can break this up and define the error using one vector $\boldsymbol{\phi}_1$:

$$\|\boldsymbol{x}\|^2 = \boldsymbol{\phi}_1^T\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{\phi}_1 + \|\boldsymbol{x}_{\text{err}}\|^2$$

Now do this for all p data points. For any single vector $\boldsymbol{\phi}_1$, we sum these together:

$$\begin{aligned}
\|\boldsymbol{x}_1\|^2 &= \boldsymbol{\phi}_1^T\boldsymbol{x}_1\boldsymbol{x}_1^T\boldsymbol{\phi}_1 + \|\boldsymbol{x}_{\text{err}}^{(1)}\|^2 \\
+\|\boldsymbol{x}_2\|^2 &= \boldsymbol{\phi}_1^T\boldsymbol{x}_2\boldsymbol{x}_2{}^T\boldsymbol{\phi}_1 + \|\boldsymbol{x}_{\text{err}}^{(2)}\|^2 \\
&\vdots \\
\|\boldsymbol{x}_p\|^2 &= \boldsymbol{\phi}_1^T\boldsymbol{x}_p\boldsymbol{x}_p{}^T\boldsymbol{\phi}_1 + \|\boldsymbol{x}_{\text{err}}^{(p)}\|^2
\end{aligned}$$

$$\sum_{k=1}^{p}\|\boldsymbol{x}_k\|^2 = \boldsymbol{\phi}_1^T\left(\sum_{k=1}^{p}\boldsymbol{x}_k\boldsymbol{x}_k{}^T\right)\boldsymbol{\phi}_1 + \sum_{k=1}^{p}\|\boldsymbol{x}_{\text{err}}^{(k)}\|^2$$

We can multiply everything by $1/(p-1)$ to make things work. That is,

$$\frac{1}{p-1}\sum_{k=1}^{p}\|\boldsymbol{x}_k\|^2 = \boldsymbol{\phi}_1^T C\boldsymbol{\phi}_1 + \frac{1}{p-1}\sum_{k=1}^{p}\|\boldsymbol{x}_{\text{err}}^{(k)}\|^2. \tag{5.1}$$

In light of this equation, let us now define an error function using an arbitrary orthonormal basis, $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n$. The error we get when using a one dimensional representation of our data is given by

$$E(\boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_n) = \frac{1}{p-1} \sum_{n=1}^{p} \|\boldsymbol{x}_{\text{err}}^{(n)}\|^2.$$

Notice that the left side of Equation 5.1 is constant (it is the sum of the magnitudes of all the known data). Therefore, **minimizing** the error function (the second term) is equivalent to **maximizing** the first term, $\boldsymbol{\phi}_1^T C \boldsymbol{\phi}_1$.

Here now is our algorithm for find the "best" basis:

1. Find the unit vector $\boldsymbol{\phi}_1$ so that $\boldsymbol{\phi}_1^T C \boldsymbol{\phi}_1$ is maximized.
2. We "project out" this vector so that the i^{th} data point now becomes:
$$\underline{\boldsymbol{x}}^{(i)} = \boldsymbol{x}^{(i)} - \text{Proj}_{\boldsymbol{\phi}_1}(\boldsymbol{x}^{(i)})$$
3. Re-compute C.
4. Repeat from Step 1 until we have enough basis vectors.

In practice, we will not need to do this- there is an easier way!

5.2 The Best Basis and the Eigenvectors

We have shown that finding the best basis reduces to maximizing the quantity:

$$\max_{\boldsymbol{\phi} \neq \mathbf{0}} \frac{\boldsymbol{\phi}^T C \boldsymbol{\phi}}{\boldsymbol{\phi}^T \boldsymbol{\phi}}$$

where we divide by the magnitude (squared) to enforce the fact that we want a unit vector, and we want to stay away from the zero vector.

We know that the eigenvectors of the covariance matrix form an orthonormal basis for $\mathbb{R}^n$, so we can write any vector as a linear combination of them:

$$\boldsymbol{\phi} = c_1 \boldsymbol{v}_1 + c_2 \boldsymbol{v}_2 + \cdots + c_n \boldsymbol{v}_n = V \boldsymbol{c}$$

Using the eigenvector-eigenvalue factorization $C = VDV^T$, we can now write the numerator as:

$$\boldsymbol{\phi}^T C \boldsymbol{\phi} = (V\boldsymbol{c})^T (VDV^T)(V\boldsymbol{c}) = \boldsymbol{c}^T (V^T V) D (V^T V) \boldsymbol{c} = \boldsymbol{c}^T D \boldsymbol{c} = c_1^2 \lambda_1 + c_2^2 \lambda_2 + \cdots + c_n^2 \lambda_n$$

Similarly, the denominator is:

$$\boldsymbol{\phi}^T \boldsymbol{\phi} = c_1^2 + c_2^2 + \cdots + c_n^2$$

Let's look at the coefficients of our expansion now. For λ_i, the coefficient in front is

$$\rho_i = \frac{c_i^2}{c_1^2 + c_2^2 + \cdots + c_n^2}$$

where $\rho_i \geq 0$ and $\sum_{i=1}^{n} \rho_i = 1$ (like a probability distribution). Let's summarize where we are. We now see that maximizing $\boldsymbol{\phi}^T C \boldsymbol{\phi}$ is equivalent to choosing $\rho_1, \rho_2, \cdots, \rho_n$ so that each $\rho_i \geq 0$ and they sum to 1, to maximize the quantity:

$$\rho_1 \lambda_1 + \rho_2 \lambda_2 + \cdots + \rho_n \lambda_n$$

It is easy to see that, if the $\lambda_i \geq 0$ and are ordered from largest to smallest, then:

$$\lambda_n \leq \rho_1\lambda_1 + \rho_2\lambda_2 + \cdots + \rho_n\lambda_n \leq \lambda_1.$$

To maximize our given quantity, we set $c_1 = 1$ and the rest of the coefficients to zero. This leads us to our main conclusion. The vector $\boldsymbol{\phi}$ that maximizes the quantity $\max_{\boldsymbol{\phi}\neq\mathbf{0}} \frac{\boldsymbol{\phi}^T C\boldsymbol{\phi}}{\boldsymbol{\phi}^T\boldsymbol{\phi}}$ is given by $\boldsymbol{v}_1$, the eigenvector corresponding to the largest eigenvalue of C. This is summarized by the theorem below:

The Best Basis Theorem

Given p points in $\mathbb{R}^n$, the *best $k-$dimensional basis* is found by taking the first k eigenvectors of the covariance matrix C. Equivalently, given the data in an $n \times p$ matrix X, the best $k-$dimensional basis is found by taking the first k columns of the U, the left singular vectors of the SVD of X. Further, this is the "best" basis for $k = 1, 2, \ldots, r$, where r is the rank of X.

Speaking of Rank...

We discussed this briefly in an earlier section, but it is worth thinking about again.

It is useful to look at the rank as that number of basis vectors required to preserve some percentage of the variance in the data. From our previous section on the covariance matrix, we had a relationship between the eigenvalues of the covariance matrix, $\hat{\lambda}_i$ and the singular values of X:

$$\frac{1}{p-1}\sigma_i^2 = \hat{\lambda}_i$$

so normalizing the set of eigenvalues is equivalent to doing it to the squared singular values:

$$\lambda_i = \frac{\hat{\lambda}_i}{\sum_{j=1}^n \hat{\lambda}_j} = \frac{\frac{1}{p-1}\sigma_i^2}{\sum_{j=1}^n \frac{1}{p-1}\sigma_j^2} = \frac{\sigma_i^2}{\sum_{j=1}^n \sigma_j^2}$$

Now the λ_i are positive and sum to 1. The idea is to keep enough dimensions r so that

$$\sum_{i=1}^{r} \lambda_i \geq \tau \qquad \text{but} \qquad \sum_{i=1}^{r-1} \lambda_i < \tau.$$

In this case, we would say that it takes r dimensions to explain or encapsulate τ percent of the variance in the data.

What should τ be? This is problem dependent. In some very noisy problems, you may only want to keep $\tau \approx 0.6$, while with very little noise, you might take $\tau \approx 0.99$.

5.2.1 The Dimensionality Reduction Step

Once we have our k basis vectors, what do we do with them? First, we create our low dimensional representation of the data. Initially, the data represents p points in $\mathbb{R}^n$, and we want to reduce that to p points in $\mathbb{R}^k$- These are the coordinates of each point using our $k-$dimensional basis. That is, if $U\Sigma V^T$ is the svd of X (mean subtracted), then the k dimensional data is created by the following, where U is $n \times k$, X is $n \times p$, and the low-dimensional representation X_{coords} is $k \times p$.

$$X_{\text{coords}} = U^T X$$

Especially if $k = 2$ or $k = 3$, we can then plot the low dimensional points in the plane or in 3-d. The "reconstruction" of the data is the representation back in $\mathbb{R}^n$ using the k basis vectors.

$$X_{\text{recon}} = UX_{\text{coords}} = UU^T X$$

Remember that earlier we said that $U^T U = I$, but UU^T is the projection matrix taking data in $\mathbb{R}^n$ and projecting it into the column space of U (so $X \neq UU^T X$ unless the columns of X are already contained within the column space of U).

5.3 Connecting to Principal Components Analysis (PCA)

In PCA, the principal components are defined to be a sequence of k direction vectors, where the i^{th} vector is the direction of a line that best fits the data while being orthogonal to the first $i-1$ vectors[1].

You can see that the principal components of set of data are then equivalent to the k basis vectors we've constructed (the first k eigenvectors of the covariance matrix). While PCA and the best basis are the same, you will typically hear the language of statistics used in PCA, while we use the language of linear algebra in constructing the best basis.

5.4 Exercises

Before doing the computer problems below, you should write down (using linear algebra notation) what computations you want to make. If you have questions (especially with the coding), I'm happy to help.

1. Suppose we have p data points in $\mathbb{R}^n$. Show that the variance of the data, projected to a standard basis vector $\boldsymbol{e}_i$, returns the usual variance for the data in that dimension. (I want you to look back at the computations we made for this in the text, "Projections and the Variance").

2. Suppose we have two o.n. vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$. Given our p points in $\mathbb{R}^n$, compute the covariance between the data projected to $\boldsymbol{u}$ and the data projected to $\boldsymbol{v}$, and (i) show that the result is

$$\boldsymbol{u}^T C \boldsymbol{v}$$

   (ii) In the special case that $\boldsymbol{u}, \boldsymbol{v}$ are eigenvectors of the covariance matrix, how does this quantity simplify?

3. Suppose we have 4 points in $\mathbb{R}^3$ as organized in the matrix X (left and below), and let $\boldsymbol{\phi}_1 = (1/\sqrt{3})[1,1,1]^T$. Use a computer (Octave/Matlab, Python or R) to compute the three quantities given in the formula to the right and below. In your script, be sure you're actually computing the covariance matrix and each quantity separately.

$$X = \begin{bmatrix} 1 & 2 & -1 & 3 \\ 0 & 0 & 1 & 1 \\ -1 & 1 & 2 & 1 \end{bmatrix}, \qquad \frac{1}{p-1}\sum_{k=1}^{p} \|\boldsymbol{x}_k\|^2 = \boldsymbol{\phi}_1^T C \boldsymbol{\phi}_1 + \frac{1}{p-1}\sum_{k=1}^{p} \|\boldsymbol{x}_{\text{err}}^{(k)}\|^2.$$

4. Using the data (and vector $\boldsymbol{\phi}_1$) in the previous exercise, compuationally verify our statements: The projection of the mean is the mean of the projection, and the variance of the data projected to $\boldsymbol{\phi}_1$ is $\boldsymbol{\phi}_1^T C \boldsymbol{\phi}_1$.

5. Verify numerically that the variance of the projected data to the first best basis vector (first one) is given by the first eigenvalue of the covariance matrix. (Careful- if you use the `eig` command, the eigenvalues are not ordered).

6. Continuing with the data from Problem 3, if we retained two of the basis vectors, how much variance (as a percentage) is "explained" by them? (This refers to the discussion in the text about how to compute the rank).

7. Load the clown data, we obtain a matrix X that is 200×320. Treat this as 320 vectors in $\mathbb{R}^{200}$.

   (a) Double center the data in X (call the result X_m).

   (b) Find the best two dimensional basis for the vectors in X_m, then project the data to two dimensions and plot the result.
   (Question to think about, you don't need to answer: Did you expect a pattern or not?)

   (c) Reconstruct the data back in $\mathbb{R}^{200}$, and show the result as an image (don't add the means back in).

[1] Wikipedia, pulled March 2021