Chapter 11

Dimensionality reduction

11.1 Random Projection and Johnson – Lindenstrauss Lemma

We saw in the previous chapter that there is a concentration effect happening in high dimension, namely that length of vectors with sub-Gaussian components tend to concentrate on an annuli. This can be leveraged as an algorithm, i.e. the random projection algorithm. This works because i.i.d. vectors with i.i.d. components are essentially orthogonal, so choosing k random vectors i.i.d. we should expect to get k decently close basis vectors of the space. Here we are relying on the independence to get orthogonality but we dont normalize length, if we did that we would loose the almost orthogonality, instead we rely on the high dimension to give us vectors that have a certain length with high probability.

Theorem 11.1 (**Random Projection**). *Let v be a fixed vector in $\mathbb{R}^d$ of length 1, fix $\epsilon \in (0,1)$ and let $U_1, \dots, U_k \in \mathbb{R}^d$ be i.i.d., mean 0, being sub-Gaussian with parameter 1 in each component and having variance a^2. Consider the projection onto $(U_1, \dots, U_k)$*

$$f(v) = (U_1 \cdot v, \dots, U_k \cdot v) : \mathbb{R}^d \to \mathbb{R}^k,$$

then

$$\mathbb{P}\left(\left||f(v)| - \sqrt{k}|a||v|\right| \geq \epsilon\sqrt{k}|a||v|\right) \leq 2e^{-\frac{k\epsilon^2}{128}}.$$

Proof. Let us first assume that $|v| = 1$. Now, since each component of U_i is sub-Gaussian with parameter 1, and the components are independent, we get

$$\mathbb{E}\left[e^{sU_i \cdot v}\right] = \prod_{j=1}^{d} \mathbb{E}\left[e^{s(U_i)_j v_j}\right] \leq \prod_{j=1}^{d} e^{s^2 (v_j)^2/2} = e^{s^2 \sum_{j=1}^{d} v_j^2/2} = e^{s^2/2}.$$

That is, $f(v)$ satisfies the prerequisites for Theorem 10.20 and we get for $\beta \leq \sqrt{k}$

$$\mathbb{P}\left(\sqrt{k}|a| - \beta \leq |f(v)| \leq \sqrt{k}|a| + \beta\right) < 2e^{-\frac{\beta^2}{128}}.$$

Setting $\epsilon = \beta/\sqrt{k} \in (0,1)$ and scaling back the length of v we get the statement of the theorem. □

Perhaps not overly exciting as we only have a single data-point, however we can use the union bound to extend this theorem to multiple points

Theorem 11.2 (**Johnson Lindenstrauss**). *For any $0 < \epsilon < 1$ and any integer n, let $k > \frac{384\ln(n)}{\epsilon^2}$. For any set of n points $\{v_1, \ldots, v_n\} \in \mathbb{R}^d$ then the random projection defined in Theorem 11.1 satisfies*

$$\mathbb{P}\left((1-\epsilon)\sqrt{k}|v_i - v_j| \leq f(v_i - v_j) \leq (1+\epsilon)\sqrt{k}|v_i - v_j|\right) \geq 1 - \frac{3}{2n}$$

Proof. Since the random projection f is linear we could for each pair $v_i - v_j$ apply Theorem 11.1 and get for $a = 1$

$$\mathbb{P}\left(\left||f(v_i - v_j)| - \sqrt{k}|v_i - v_j|\right| \geq \epsilon\sqrt{k}|v_i - v_j|\right) \leq 2e^{-\frac{k\epsilon^2}{128}}.$$

There are $\binom{n}{2} < n^2/2$ pairs, so by the union bound we get

$$\mathbb{P}\left(\exists i, j : \left||f(v_i - v_j)| - \sqrt{k}|v_i - v_j|\right| \geq \epsilon\sqrt{k}|v_i - v_j|\right) \leq n^2 e^{-\frac{k\epsilon^2}{128}}.$$

Now, if we choose k such that

$$n^2 e^{-\frac{k\epsilon^2}{128}} = 1/n$$

this becomes

$$k = \frac{384\ln(n)}{\epsilon^2}.$$

□

Remark 11.3. *Note that this usually requires k to be quite large, however we are proving the probability that all distances are preserved. It is usually better if we can allow more error, see Fig. 11.1. The reason for this is that the data itself is IID and as such we can think of Theorem 11.1 as providing a p for a Bernoulli trial, but this is of course not rigorous.*

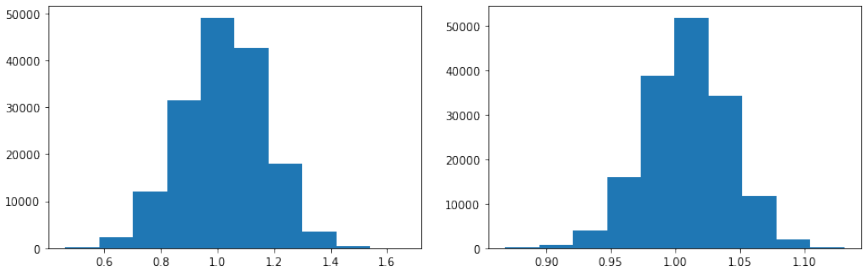


Figure 11.1: The distribution of relative error on the Olivetti faces dataset using only $k = 20$ and $k = 400$ respectively.

11.2 SVD (Singular Value Decomposition)

Something that works "better" in medium high dimension (whatever that means) is **SVD** or **Singular Value Decomposition**.

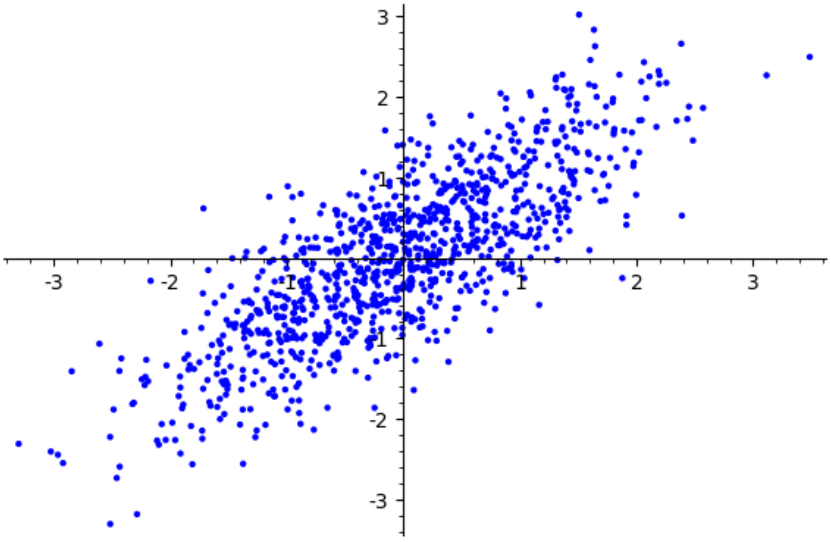


Figure 11.2: Sample data for SVD

Lets say that we wish to represent the data using a low-dimensional subspace, think of a low-dimensional plane. In the case of 2d there is only 1d planes (lines), but if you have, say 100 dimensions we could consider the best fitting 10 dimensional plane. What we mean with best fitting is that the distance from the point to its projection onto our subspace is as small as possible. Think of our 2d example above, then we would like to find the line such that orthogonal projection gives the smallest error. Just looking at the plot we would take the line $y = x$.

But how do we formulate this rigorously? Well we will solve another problem, and later see that it is the same

Remark 11.4. *Consider a line given by the unit vector v, and consider a point x then the projection of x onto v is as above given by*

$$(v \cdot x)v$$

We will now use these ideas applied to IID samples of points $\{X_1, \ldots, X_n\} \in \mathbb{R}^m$ with zero empirical mean (we have centered them). Let $v \in \mathbb{R}^m$ be a unit vector. Consider the projection of each X_i onto v but only consider the proportion i.e. $X_i \cdot v$, then define

$$Y_i = (X_i \cdot v)$$

The line with maximal empirical variance can be written as ($\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i = \frac{1}{n}\sum_{i=1}^n X \cdot v = 0$ since we assumed zero empirical mean)

$$\begin{aligned} v_1 &:= \arg\max_{\|v\|=1} \frac{1}{n}\sum_i (Y_i - \overline{Y}_n)^2 \\ &= \arg\max_{\|v\|=1} \sum_{j=1}^n |X_i \cdot v|^2. \end{aligned}$$

If we construct a matrix A of size $n \times m$ with rows X_i then we can rewrite

$$\sum_{j=1}^n |X_i \cdot v|^2 = |Av|^2$$

and our problem reduces to the linear algebra problem of given an $n \times m$ matrix A to find the direction that is most "expanded/least contracted" by A, in the following sense

$$\arg\max_{\|w\|=1} |Av|.$$

Remark 11.5. *Note, the singular vectors are not necessarily unique, in fact if v is a singular vector, then so is $-v$. We can also have ties, in that case we arbitrarily pick one. We assume that the singular vectors can be picked uniquely, for instance by requiring no ties and that we fix the sign as to make the vector unique.*

Definition 11.6. *The vector $v_1 \in \mathbb{R}^m$ of the $(n \times m)$ matrix A, defined as*

$$v_1 := \arg\max_{\|v\|=1} |Av|$$

is called the **first singular vector** *of A. The value $\sigma_1(A)$ defined as*

$$\sigma_1(A) := |Av_1|$$

is called the **first singular value** *of A.*

Now that we have defined the first singular vector, we can define the second singular vector. This is simply a vector that is orthogonal to v_1 again solving our maximum problem, i.e.

$$v_2 := \arg \max_{\|v\|=1, v \perp v_1} |Av|.$$

We can interpret this as follows, consider the plane given by the first singular vector v_1 as the normal, then we can consider our new problem by finding the vector v that maximizes $|(P_1 A)v|$ where

$$PA = \begin{bmatrix} P_1 X_1 \\ P_1 X_2 \\ \vdots \\ P_1 X_n \end{bmatrix}, \quad P_1 x = x - (x \cdot v_1) v_1.$$

where P_1 is the projection of a vector onto the plane $v_1 \cdot x = 0$. See Fig. 11.3 for the result of the projection.

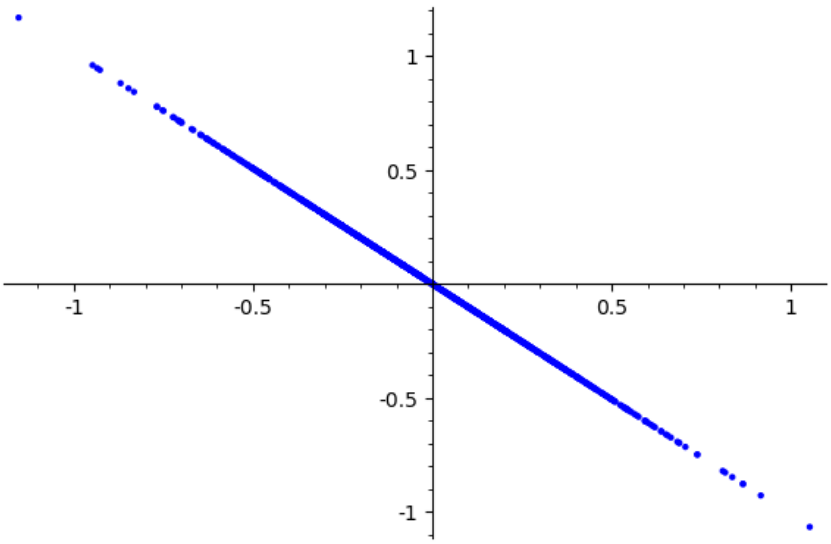


Figure 11.3: The data from Fig. 11.2 projected onto the normal of the plane defined by v_1.

This can be extended, all the way until we have m vectors. That is, there are m singular vectors. To connect this to something which you have already seen in linear algebra, note that

$$\arg \max_{\|v\|=1} |Av| = \arg \max_{\|v\|=1} |Av|^2 = \arg \max_{\|v\|=1} \langle Av, Av \rangle = \arg \max_{\|v\|=1} \langle A^T Av, v \rangle$$

If we let $(v_1, \ldots, v_m)$ be the eigenvectors (all orthogonal) and $\lambda_1, \ldots, \lambda_m$ be the eigenvalues of $A^T A$ (all positive and ordered decreasingly) then we can write

$$v = \sum_{i=1}^{m} a_i v_i$$

which allows us to write

$$\begin{aligned}\langle A^T A v, v\rangle &= \left\langle A^T A \left(\sum_{i=1}^{m} a_i v_i\right), \sum_{i=1}^{m} a_i v_i \right\rangle \\ &= \sum_{i=1}^{m} \left\langle \lambda_i a_i v_i, \sum_{i=1}^{m} a_i v_i \right\rangle = \sum_{i=1}^{m} \lambda_i a_i^2\end{aligned}$$

now, since $1 = \|v\| = \sqrt{a_1^2 + \ldots + a_m^2}$, the above is maximized if $a_1 = 1$ and all other $a_i = 0$, since λ_1 is the largest eigenvalue. We have now proved

Lemma 11.7. *Let A be an $(n \times m)$ matrix, and let v_1 be the first singular vector of A and let $\sigma_1(A)$ be the first singular value (with $\sigma_2(A) < \sigma_1(A)$), then v_1 is an eigenvector of the $m \times m$ matrix $A^T A$, and*

$$\max_{\|v\|=1} |Av| = |Av_1| = \sqrt{\lambda_1} = \sigma_1(A)$$

where λ_1 is the first eigenvalue of $A^T A$.

Remark 11.8. *The problem which can occur in the above is that for multiple eigenvalues we have to choose one of them and identify that with the singular vector, but this is just up to a permutation of indices. Secondly, the above works for any singular value. I.e. every singular vector is an eigenvector and every singular value is the square root of an eigen-value.*

Remark 11.9. *In the context where A is constructed from our IID vectors $X_1, \ldots, X_n$, we see that σ_1 is the standard deviation in the direction of the first singular vector. Furthermore, the matrix $A^T A$ will then be the empirical covariance matrix. I.e. we are looking at the eigenvectors and eigenvalues of the empirical covariance matrix.*

Theorem 11.10 (**Greedy Algorithm**). *Let A be an $n \times d$ matrix with singular vectors $v_1, ..., v_r$. For $1 \leq k \leq r$, let V_k be the subspace spanned by $v_1, \ldots, v_k$. For each k, V_k is the best fit k-dimensional subspace for A.*

Here, V_k is defined as

$$V_k = \{\alpha_1 v_1 + \ldots + \alpha_k v_k : (\alpha_1, \ldots, \alpha_k) \in \mathbb{R}^k\} =: span(\{v_1, \ldots, v_k\}).$$

What do we mean by best fit? Let $\tilde{V}_k$ be another k-dimensional subspace consider the distance of a point p to the k-dimensional subspace $\tilde{V}_k$, such a space is spanned by an orthonormal basis $\tilde{v}_1, \ldots, \tilde{v}_k$, the distance from p to $\tilde{V}_k$ can be seen to be

$$\|p - proj_{\tilde{V}_k} p\| = \|p - \sum_{i=1}^{k} (\tilde{v}_i \cdot p)\tilde{v}_i\|$$

We mean that V_k is the k-dimensional subspace that minimizes

$$\sum_{i=1}^{n} \|X_i - proj_{\tilde{V}_k} X_i\|^2$$

But we can use the Pythagorean theorem to get

$$\sum_{i=1}^{n} \left(\|proj_{\tilde{V}_k} X_i\|^2 + \|X_i - proj_{\tilde{V}_k} X_i\|^2 \right) = \sum_{i=1}^{n} \|X_i\|^2$$

and thus we can get

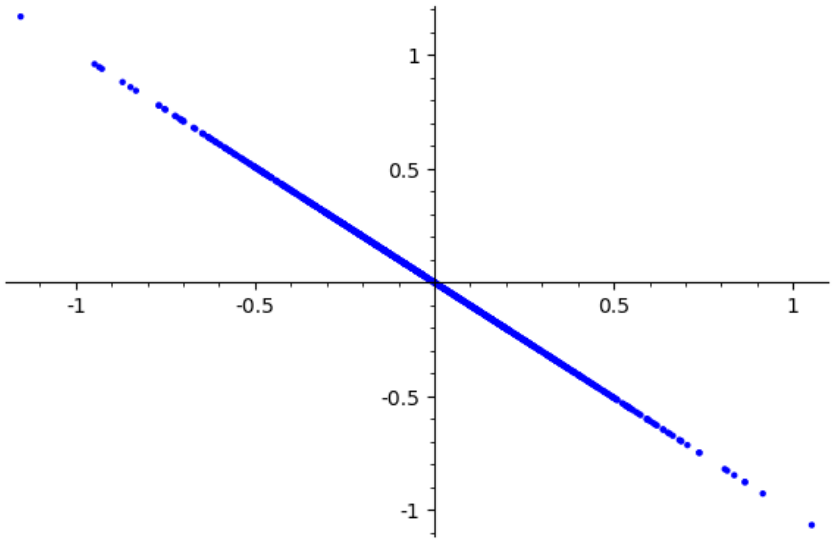
$$\sum_{i=1}^{n} \left(\|X_i\|^2 - \|proj_{\tilde{V}_k} X_i\|^2 \right) = \sum_{i=1}^{n} \|X_i - proj_{\tilde{V}_k} X_i\|^2$$

From the above we see that the best fitting subspace is the subspace that maximizes the "variance" in the sense that we have seen. The point I am making is that we can rephrase the theorem as saying that finding $v_1, \ldots, v_k$ in a greedy way by maximizing the variance is the same as directly minimizing the variance of the deviation from the subspace. This thus answers our question in the beginning of the section.

If we run the greedy algorithm we get the following on the data plotted above.

```
[-0.71191709 -0.70226352] 43.587923587503624
```

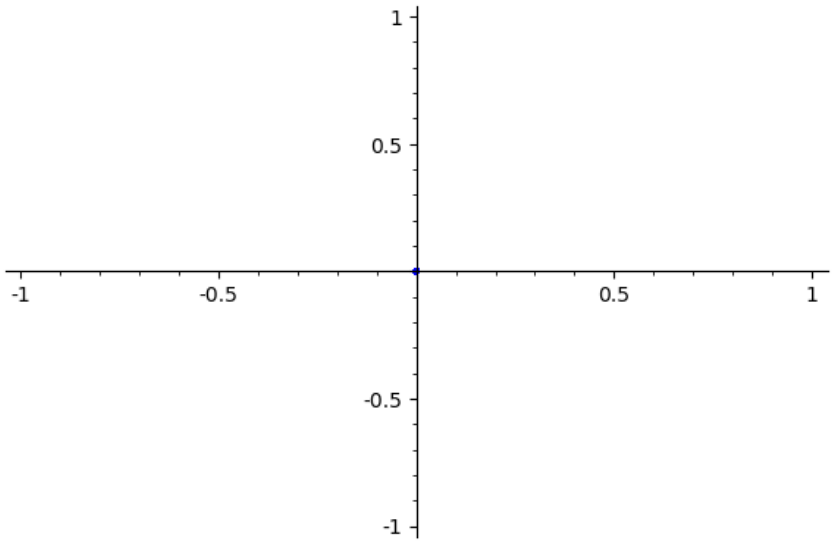
As we can see this is pretty much identical to the vector in the 45 degree direction. To find the second singular vector, we simply project our data onto the plane spanned by having v_1 as the normal.



We then get the following

```
[ 0.70226351 -0.7119171 ] 14.365618434588386
```

Its clear which direction this is headed. Let us also look at what happens when we project the data onto the plane with normal v_2.



Exercise 11.11. *What have we done? Two projections in a row? What is the projection of a projection?*

It should be clear from the definition and Theorem 11.10 that $proj_{V_m} A = A$. That is, if we use all possible singular vectors, then we can represent the data from A as points in V_m. That is any row of A can be written as a linear combination of all the singular vectors.

Singular Value Decomposition of a Matrix

Remember that we said that if we compute m singular vectors of the $n \times m$ dimensional matrix A, then

$$proj_{V_m} A = A$$

this implies that we can write each row in A as $X_i = \sum_{j=1}^{m} (X_i \cdot v_j) v_j$ which we can now rewrite as

$$A = \sum_{j=1}^{m} A v_j v_j^T$$

denoting $u_i := \frac{Av_i}{\sigma_i}$ we see that the above expression becomes

$$A = \sum_{j=1}^{m} \sigma_j u_j v_j^T \tag{11.1}$$

This is the singular value decomposition. I.e. we have decomposed A into a sum of matrices, that is $u_j v_j^T$ is $n \times m$ matrices.

Rewriting the above equation in matrix format we get

$$A = UDV^T$$

where U is the matrix with $u_1, \ldots$ as the columns, D is a diagonal matrix with σ_i as the diagonal and V is the matrix with v_i as columns of V.

Definition 11.12. *The vectors u_i are called the* **left singular vectors**.

11.2.1 The power method

Another way to prove Lemma 11.7 is to consider the matrix $A^T A$ using our decomposition above to get

$$A^T A = (UDV^T)^T (UDV^T) = (VDU^T UDV^T) = VD^2 V^T$$

since $U^T U = I$ which comes from the fact that the columns are orthonormal.

Exercise 11.13. *Prove that the left singular vectors are orthonormal.*

This means that for any column v_i in V

$$A^T A v_i = V D^2 V^T v_i = \sigma_i^2 v_i$$

so we see that v_i is the i:th eigenvector of $A^T A$ with eigenvalue σ_i^2. We can thus find the singular vectors by trying to find the eigenvectors of $A^T A$.

How do we find the eigenvectors of $B = A^T A$? Well first note that

$$B^k = (V D^2 V^T)^k = (V D^{2k} V^T)$$

by the same argument as above, i.e. $V^T V = I$. Thus we see that if $\sigma_1 > \sigma_2$ then if we let k be large enough then

$$B^k \approx (\sigma_1)^{2k} v_1 v_1^T.$$

11.3 PCA

What is PCA, well basically it is a coordinate transformation from the original coordinates to the coordinate system given by the singular vectors. Since V is orthonormal it is as simple as a product, i.e.

$$A = UDV^T$$

Recall that each row in A is a data point i.e. an m dimensional vector and that V is an orthonormal basis, as such we project each point in A onto each basis vector from V by using dot products, as in $(X_i \cdot v_i)v_i$, the coordinate in the basis is just $X_i \cdot v_i$, and as such we get

$$PCA(A) = AV = UDV^T V = UD$$

Remark 11.14. *Warning: In the beginning of this section we assumed that our data had empirical mean zero. Thus in order to use this we first have to center the data.*

11.4 SVD in Action

This is all cool and such, but what can you do with it?

Singular value decomposition can be used in the following ways

11.4.1 Factor Analysis

- Studying underlying factors. The famous `g factor`: proposed by Spearman (Spearman correlation), to describe "general intelligence" as a singular vector based on data about IQ, Math ability and other cognitive tests. This is also called Factor analysis.
- Compressing a representation of data, as a dimensional reduction technique. This is similar to the rank k approximation idea.

11.4.2 Example on compressing data

Lets consider the Mnist dataset, which is handwritten digits from 0 to 9. These are represented as 8×8 pixel images and will be put together as a single array of length 64. As such we have points in $\mathbb{R}^d$ with $d = 64$. The number of points is 1797. If we assemble all these images into a matrix as before, where each row is a datapoint (image) we get a matrix A of shape 1797×64. Recall from (11.1) we have that A is the sum of m matrices of shape $n \times m$, we will now sum this from 1 to 10 instead and consider

$$A_k := \sum_{i=1}^{k} \sigma_j u_j v_j^T$$

for $k = 10$. That is, we are using 10 singular vectors to represent the digits, in Fig. 11.4 you can see 10 uncompressed sample images and in Fig. 11.5 you can see the same 10 samples but compressed. What do we mean, we mean that if X_i is an image, it will be row i of A, the compressed image will be row i of A_{10}.

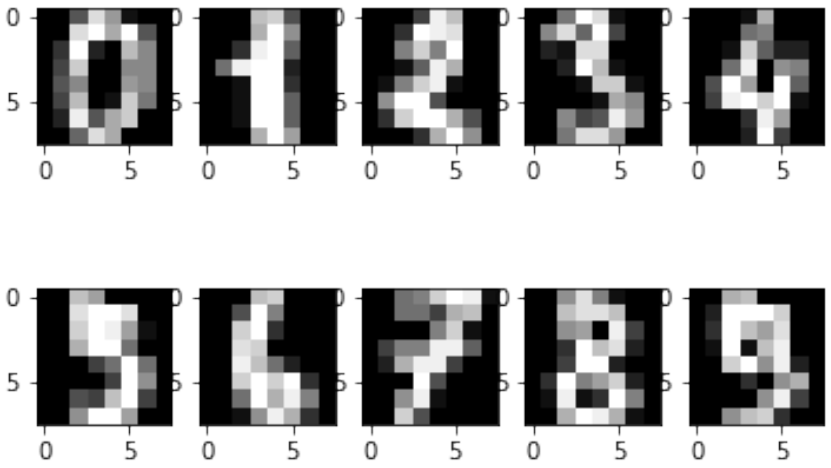


Figure 11.4: 10 sample images from Mnist

```
Number of data_points: 1797, number of features: 64,␣
↪Number of components: 10
```

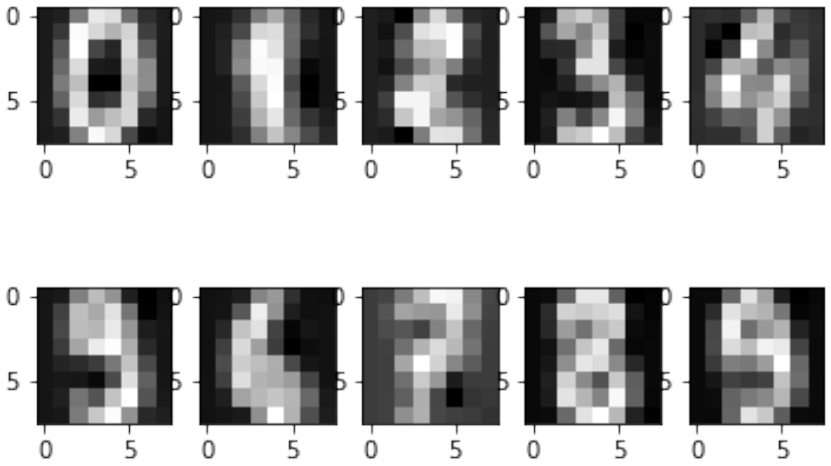


Figure 11.5: The data from Fig. 11.4 projected onto the plane defined by the first 10 singular vectors.

What we can see is that even with only 10 components we were able to fairly well represent the digits, although it is clear that some are not so easy.

Reconstruction error

The reconstruction error is defined as the error we make in the compression, i.e. the square distance between the real image and the target image,

$$\text{Reconst} := \sqrt{\sum_{i=1}^{n} \|(A)_i - (A_k)_i\|^2}$$

For those who know linear algebra this is nothing else than the Frobenious norm of $A - A_k$. Using the decomposition (11.1) we get that

$$A - A_k = \sum_{j=k+1}^{m} \sigma_j u_j v_j^T$$

and the norm of this is simply $\sqrt{\sum_{j=k+1}^{m} \sigma_j^2}$. As such the sum of squares of the remaining singular values are giving us the reconstruction error.

Explained variance

Explained variance is how much percentage of the total variance is captured by our singular vectors. Remember the interpretation of the singular values as the standard deviation, as such the variance explained of the first k components is just the sum of the singular values squared and divided by the total variance.

11.4.3 Anomaly detection and reconstruction error

The approach taken in Section 11.4.2 can be used for a rudimentary form of anomaly detection, which incidentally works really well.

The point is here is that we compress data into the matrix A_k, we estimate the distribution function for $\|(A)_i - (A_k)_i\|$ using the samples and then use this to select quantiles that we will use for detection of an anomaly.

11.5 Theoretical analysis

The PCA components are eigenvector of the empirical covariance matrix. Namely, let $Z = (X_1, \ldots, X_d) \sim F_Z$ and consider an i.i.d. sequence of $Z_1, \ldots, Z_n$. Covariance matrix is

$$\mathbb{E}\left[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])^T\right]$$

assuming that Z has mean zero, lets consider

$$\mathbb{E}[ZZ^T] = (\mathbb{E}[X_i X_j])_{i,j}$$

there are $d^2/2$ such values. Now the empirical covariance matrix is that we use the empirical mean to estimates each component of the matrix, i.e.

$$\hat{\Sigma}_{i,j} = \frac{1}{n}\sum_{k=1}^{n}(Z_k)_i(Z_k)_j$$

if now each component of Z_k is sub-Gaussian then we can use concentration to get something like

$$\mathbb{P}(|\hat{\Sigma}_{i,j} - (\mathbb{E}[X_i X_j])_{i,j}| > \epsilon) < e^{-c\epsilon n}$$

using the union bound we can thus get

$$\mathbb{P}(\max_{i,j}|\hat{\Sigma}_{i,j} - (\mathbb{E}[X_i X_j])_{i,j}| > \epsilon) < \frac{d^2}{2}e^{-c\epsilon n}$$

The d^2 in the estimate is however quite suboptimal and there is an improvement over the above, which follows from the so-called **Matrix Bernstein inequality**.

Theorem 11.15. *Let $X_1, \ldots, X_n$ be centred i.i.d, random vectors in $\mathbb{R}^d$. Suppose that for all i, $\mathrm{Var}(X_i) = \Sigma$ and $\mathbb{P}(\|X_i\|_2 \leq \sqrt{C}) = 1$ for some C. Then for all $\epsilon > 0$*

$$\mathbb{P}\left[\left\|\hat{\Sigma}_n - \Sigma\right\| > \epsilon\right] \leq 2de^{-\frac{n\epsilon^2}{2C(C+2\epsilon/3)}}$$

Remark 11.16. *The notation $\|\Sigma\|$ for matrices, denotes the operator norm.*

This theorem tells us that with high probability the estimated covariance matrix will be close to the true covariance Σ if we have many observations. However, the PCA method relied on computing the eigen-values and eigen-vectors of $\hat{\Sigma}_n$. Closeness in the matrix norm allows us to say something about the closeness of the eigen-values

Theorem 11.17 (Weyls theorem)**.** *Let $\hat{\Sigma} = \Sigma + E$, where Σ and E are symmetric matrices. Let λ_i and $\hat{\lambda}_i$ be the i:th eigen-values of Σ and $\hat{\Sigma}$ respectively. Then*

$$\max_{i=1,\dots,d} |\hat{\lambda}_i - \lambda_i| \leq \|E\|.$$

This is all well and good, but estimating the eigen-vectors is a difficult problem.

Example 11.18. *Consider $\Sigma = \begin{bmatrix} 1, & 0 \\ 0, & 1 \end{bmatrix}$ and $E = \begin{bmatrix} 0, & \epsilon \\ \epsilon, & 0 \end{bmatrix}$. The eigenvalues of Σ are 1 and 1. The eigen-vectors of Σ are $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$. The eigen-values of $\hat{\Sigma} := \Sigma + E$ is $1 + \epsilon$ and $1 - \epsilon$. However, for any ϵ, the eigenvectors of $\hat{\Sigma}$ are $\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.*

The problem in the above example is the closeness of the eigen-values for $\hat{\Sigma}$, since Σ has a double eigen-value. This poses problems as it is an unstable problem and we have no hope. What we can say though, is that if the eigen-values are only simple, then we can expect stability. We will not cover that in this course, but if you want to dig deeper, check-out [WW].

11.6 Reconstruction error

Introduce the class

$$\mathcal{P}_k = \{\Pi : \mathbb{R}^d \to \mathbb{R}^d \mid \Pi \text{ is an orthogonal projection of rank } k\}.$$

Consider a $\mathbb{R}^d$ valued random variable $X \in L^2(\mathbb{P})$. Define the loss function $L(X, \Pi(X)) = \|X - \Pi(X)\|_2^2$, for $X \in \mathbb{R}^d$, then define the reconstruction error of the projection operator Π as

$$\mathcal{R}(\Pi) = \mathbb{E}\left[L(Z, \Pi(Z))\right].$$

The minimizer of the risk Π_k^* is defined as

$$\Pi_k^* = \arg\min_{\Pi \in \mathcal{P}_k} \mathcal{R}(\Pi).$$

As we have seen above with singular value decomposition etc. we have that Π_k^* is the projection onto the first k eigen-vectors of the covariance matrix Σ.

The empirical minimization problem is

$$\hat{\Pi}_k^* = \arg\min_{\Pi \in \mathcal{P}_k} \frac{1}{n} \sum_{i=1}^{n} L(X_i, \Pi(X_i)) = \arg\min_{\Pi \in \mathcal{P}_k} \frac{1}{n} \sum_{i=1}^{n} \|X_i - \Pi(X_i)\|_2^2.$$

The excess risk is defined as

$$\mathcal{E}_k := \mathcal{R}(\hat{\Pi}_k^*) - \mathcal{R}(\Pi_k^*)$$

as in Chapter 8 the goal is to bound the excess risk with high probability.

We have the following estimate

Lemma 11.19. *In the setting above, if we define $\hat{\Sigma}$ the empirical covariance matrix, then*

$$\mathcal{E}_k \leq \sqrt{2k} \|\Sigma - \hat{\Sigma}\|_2$$

Proof. See [ReWa, Proposition 2.2]. □

Thus, we have from Theorem 11.15

Theorem 11.20. *Let $X_1, \ldots, X_n$ be centred i.i.d, random vectors in $\mathbb{R}^d$. Suppose that for all i, $\mathrm{Var}(X_i) = \Sigma$ and $\mathbb{P}(\|X_i\|_2 \leq \sqrt{C}) = 1$ for some C. Then for all $\epsilon > 0$*

$$\mathbb{P}\left[\mathcal{E}_k > \epsilon\right] \leq \mathbb{P}\left[\left\|\hat{\Sigma}_n - \Sigma\right\| > \epsilon/\sqrt{2k}\right] \leq 2de^{-\frac{n\epsilon^2}{4C(C+2\epsilon/3)k}}$$

11.7 Bibliography

The first part concerning SVD is losely built on [BlHo]. For the Bernstein inequality Theorem 11.15 see [WW, Corollary 6.20]. If you want to dig deeper into reconstruction errors, see [ReWa].