23 Dimensionality Reduction

Dimensionality reduction is the process of taking data in a high dimensional space and mapping it into a new space whose dimensionality is much smaller. This process is closely related to the concept of (lossy) compression in information theory. There are several reasons to reduce the dimensionality of the data. First, high dimensional data impose computational challenges. Moreover, in some situations high dimensionality might lead to poor generalization abilities of the learning algorithm (for example, in Nearest Neighbor classifiers the sample complexity increases exponentially with the dimension—see Chapter 19). Finally, dimensionality reduction can be used for interpretability of the data, for finding meaningful structure of the data, and for illustration purposes.

In this chapter we describe popular methods for dimensionality reduction. In those methods, the reduction is performed by applying a linear transformation to the original data. That is, if the original data is in $\mathbb{R}^d$ and we want to embed it into $\mathbb{R}^n$ $(n < d)$ then we would like to find a matrix $W \in \mathbb{R}^{n,d}$ that induces the mapping $\mathbf{x} \mapsto W\mathbf{x}$. A natural criterion for choosing W is in a way that will enable a reasonable recovery of the original $\mathbf{x}$. It is not hard to show that in general, exact recovery of $\mathbf{x}$ from $W\mathbf{x}$ is impossible (see Exercise 1).

The first method we describe is called Principal Component Analysis (PCA). In PCA, both the compression and the recovery are performed by linear transformations and the method finds the linear transformations for which the differences between the recovered vectors and the original vectors are minimal in the least squared sense.

Next, we describe dimensionality reduction using random matrices W. We derive an important lemma, often called the "Johnson-Lindenstrauss lemma," which analyzes the distortion caused by such a random dimensionality reduction technique.

Last, we show how one can reduce the dimension of all *sparse* vectors using again a random matrix. This process is known as Compressed Sensing. In this case, the recovery process is nonlinear but can still be implemented efficiently using linear programming.

We conclude by underscoring the underlying "prior assumptions" behind PCA and compressed sensing, which can help us understand the merits and pitfalls of the two methods.

23.1 Principal Component Analysis (PCA)

Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be m vectors in $\mathbb{R}^d$. We would like to reduce the dimensionality of these vectors using a linear transformation. A matrix $W \in \mathbb{R}^{n,d}$, where $n < d$, induces a mapping $\mathbf{x} \mapsto W\mathbf{x}$, where $W\mathbf{x} \in \mathbb{R}^n$ is the lower dimensionality representation of $\mathbf{x}$. Then, a second matrix $U \in \mathbb{R}^{d,n}$ can be used to (approximately) recover each original vector $\mathbf{x}$ from its compressed version. That is, for a compressed vector $\mathbf{y} = W\mathbf{x}$, where $\mathbf{y}$ is in the low dimensional space $\mathbb{R}^n$, we can construct $\tilde{\mathbf{x}} = U\mathbf{y}$, so that $\tilde{\mathbf{x}}$ is the recovered version of $\mathbf{x}$ and resides in the original high dimensional space $\mathbb{R}^d$.

In PCA, we find the compression matrix W and the recovering matrix U so that the total squared distance between the original and recovered vectors is minimal; namely, we aim at solving the problem

$$\operatorname*{argmin}_{W \in \mathbb{R}^{n,d}, U \in \mathbb{R}^{d,n}} \sum_{i=1}^{m} \|\mathbf{x}_i - UW\mathbf{x}_i\|_2^2. \tag{23.1}$$

To solve this problem we first show that the optimal solution takes a specific form.

LEMMA 23.1 *Let (U, W) be a solution to Equation (23.1). Then the columns of U are orthonormal (namely, $U^\top U$ is the identity matrix of $\mathbb{R}^n$) and $W = U^\top$.*

Proof Fix any U, W and consider the mapping $\mathbf{x} \mapsto UW\mathbf{x}$. The range of this mapping, $R = \{UW\mathbf{x} : \mathbf{x} \in \mathbb{R}^d\}$, is an n dimensional linear subspace of $\mathbb{R}^d$. Let $V \in \mathbb{R}^{d,n}$ be a matrix whose columns form an orthonormal basis of this subspace, namely, the range of V is R and $V^\top V = I$. Therefore, each vector in R can be written as $V\mathbf{y}$ where $\mathbf{y} \in \mathbb{R}^n$. For every $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^n$ we have

$$\|\mathbf{x} - V\mathbf{y}\|_2^2 \;=\; \|\mathbf{x}\|^2 + \mathbf{y}^\top V^\top V\mathbf{y} - 2\mathbf{y}^\top V^\top \mathbf{x} \;=\; \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\mathbf{y}^\top (V^\top \mathbf{x}),$$

where we used the fact that $V^\top V$ is the identity matrix of $\mathbb{R}^n$. Minimizing the preceding expression with respect to $\mathbf{y}$ by comparing the gradient with respect to $\mathbf{y}$ to zero gives that $\mathbf{y} = V^\top \mathbf{x}$. Therefore, for each $\mathbf{x}$ we have that

$$VV^\top \mathbf{x} = \operatorname*{argmin}_{\tilde{\mathbf{x}} \in R} \|\mathbf{x} - \tilde{\mathbf{x}}\|_2^2.$$

In particular this holds for $\mathbf{x}_1, \ldots, \mathbf{x}_m$ and therefore we can replace U, W by $V, V^\top$ and by that do not increase the objective

$$\sum_{i=1}^{m} \|\mathbf{x}_i - UW\mathbf{x}_i\|_2^2 \;\geq\; \sum_{i=1}^{m} \|\mathbf{x}_i - VV^\top \mathbf{x}_i\|_2^2.$$

Since this holds for every U, W the proof of the lemma follows. □

On the basis of the preceding lemma, we can rewrite the optimization problem given in Equation (23.1) as follows:

$$\operatorname*{argmin}_{U \in \mathbb{R}^{d,n} : U^\top U = I} \sum_{i=1}^{m} \|\mathbf{x}_i - UU^\top \mathbf{x}_i\|_2^2. \tag{23.2}$$

We further simplify the optimization problem by using the following elementary algebraic manipulations. For every $\mathbf{x} \in \mathbb{R}^d$ and a matrix $U \in \mathbb{R}^{d,n}$ such that $U^\top U = I$ we have

$$\begin{aligned}\|\mathbf{x} - UU^\top \mathbf{x}\|^2 &= \|\mathbf{x}\|^2 - 2\mathbf{x}^\top UU^\top \mathbf{x} + \mathbf{x}^\top UU^\top UU^\top \mathbf{x} \\ &= \|\mathbf{x}\|^2 - \mathbf{x}^\top UU^\top \mathbf{x} \\ &= \|\mathbf{x}\|^2 - \operatorname{trace}(U^\top \mathbf{x}\mathbf{x}^\top U), \end{aligned} \tag{23.3}$$

where the trace of a matrix is the sum of its diagonal entries. Since the trace is a linear operator, this allows us to rewrite Equation (23.2) as follows:

$$\operatorname*{argmax}_{U \in \mathbb{R}^{d,n} : U^\top U = I} \operatorname{trace}\left(U^\top \sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i^\top U \right). \tag{23.4}$$

Let $A = \sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i^\top$. The matrix A is symmetric and therefore it can be written using its spectral decomposition as $A = VDV^\top$, where D is diagonal and $V^\top V = VV^\top = I$. Here, the elements on the diagonal of D are the eigenvalues of A and the columns of V are the corresponding eigenvectors. We assume without loss of generality that $D_{1,1} \geq D_{2,2} \geq \cdots \geq D_{d,d}$. Since A is positive semidefinite it also holds that $D_{d,d} \geq 0$. We claim that the solution to Equation (23.4) is the matrix U whose columns are the n eigenvectors of A corresponding to the largest n eigenvalues.

THEOREM 23.2 *Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be arbitrary vectors in $\mathbb{R}^d$, let $A = \sum_{i=1}^m \mathbf{x}_i \mathbf{x}_i^\top$, and let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be n eigenvectors of the matrix A corresponding to the largest n eigenvalues of A. Then, the solution to the PCA optimization problem given in Equation (23.1) is to set U to be the matrix whose columns are $\mathbf{u}_1, \ldots, \mathbf{u}_n$ and to set $W = U^\top$.*

Proof Let $VDV^\top$ be the spectral decomposition of A. Fix some matrix $U \in \mathbb{R}^{d,n}$ with orthonormal columns and let $B = V^\top U$. Then, $VB = VV^\top U = U$. It follows that

$$U^\top AU = B^\top V^\top VDV^\top VB = B^\top DB,$$

and therefore

$$\operatorname{trace}(U^\top AU) = \sum_{j=1}^d D_{j,j} \sum_{i=1}^n B_{j,i}^2.$$

Note that $B^\top B = U^\top VV^\top U = U^\top U = I$. Therefore, the columns of B are also orthonormal, which implies that $\sum_{j=1}^d \sum_{i=1}^n B_{j,i}^2 = n$. In addition, let $\tilde{B} \in \mathbb{R}^{d,d}$ be a matrix such that its first n columns are the columns of B and in addition $\tilde{B}^\top \tilde{B} = I$. Then, for every j we have $\sum_{i=1}^d \tilde{B}_{j,i}^2 = 1$, which implies that $\sum_{i=1}^n B_{j,i}^2 \leq 1$. It follows that:

$$\operatorname{trace}(U^\top AU) \leq \max_{\boldsymbol{\beta} \in [0,1]^d \,:\, \|\boldsymbol{\beta}\|_1 \leq n} \sum_{j=1}^d D_{j,j} \beta_j \ .$$

It is not hard to verify (see Exercise 2) that the right-hand side equals to $\sum_{j=1}^{n} D_{j,j}$. We have therefore shown that for every matrix $U \in \mathbb{R}^{d,n}$ with orthonormal columns it holds that $\text{trace}(U^\top AU) \leq \sum_{j=1}^{n} D_{j,j}$. On the other hand, if we set U to be the matrix whose columns are the n leading eigenvectors of A we obtain that $\text{trace}(U^\top AU) = \sum_{j=1}^{n} D_{j,j}$, and this concludes our proof. □

Remark 23.1 The proof of Theorem 23.2 also tells us that the value of the objective of Equation (23.4) is $\sum_{i=1}^{n} D_{i,i}$. Combining this with Equation (23.3) and noting that $\sum_{i=1}^{m} \|\mathbf{x}_i\|^2 = \text{trace}(A) = \sum_{i=1}^{d} D_{i,i}$ we obtain that the optimal objective value of Equation (23.1) is $\sum_{i=n+1}^{d} D_{i,i}$.

Remark 23.2 It is a common practice to "center" the examples before applying PCA. That is, we first calculate $\boldsymbol{\mu} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i$ and then apply PCA on the vectors $(\mathbf{x}_1 - \boldsymbol{\mu}), \ldots, (\mathbf{x}_m - \boldsymbol{\mu})$. This is also related to the interpretation of PCA as variance maximization (see Exercise 4).

23.1.1 A More Efficient Solution for the Case $d \gg m$

In some situations the original dimensionality of the data is much larger than the number of examples m. The computational complexity of calculating the PCA solution as described previously is $O(d^3)$ (for calculating eigenvalues of A) plus $O(md^2)$ (for constructing the matrix A). We now show a simple trick that enables us to calculate the PCA solution more efficiently when $d \gg m$.

Recall that the matrix A is defined to be $\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i^\top$. It is convenient to rewrite $A = X^\top X$ where $X \in \mathbb{R}^{m,d}$ is a matrix whose ith row is $\mathbf{x}_i^\top$. Consider the matrix $B = XX^\top$. That is, $B \in \mathbb{R}^{m,m}$ is the matrix whose i, j element equals $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$. Suppose that $\mathbf{u}$ is an eigenvector of B: That is, $B\mathbf{u} = \lambda\mathbf{u}$ for some $\lambda \in \mathbb{R}$. Multiplying the equality by $X^\top$ and using the definition of B we obtain $X^\top XX^\top \mathbf{u} = \lambda X^\top \mathbf{u}$. But, using the definition of A, we get that $A(X^\top \mathbf{u}) = \lambda(X^\top \mathbf{u})$. Thus, $\frac{X^\top \mathbf{u}}{\|X^\top \mathbf{u}\|}$ is an eigenvector of A with eigenvalue of λ.

We can therefore calculate the PCA solution by calculating the eigenvalues of B instead of A. The complexity is $O(m^3)$ (for calculating eigenvalues of B) and $m^2 d$ (for constructing the matrix B).

Remark 23.3 The previous discussion also implies that to calculate the PCA solution we only need to know how to calculate inner products between vectors. This enables us to calculate PCA implicitly even when d is very large (or even infinite) using kernels, which yields the *kernel PCA* algorithm.

23.1.2 Implementation and Demonstration

A pseudocode of PCA is given in the following.

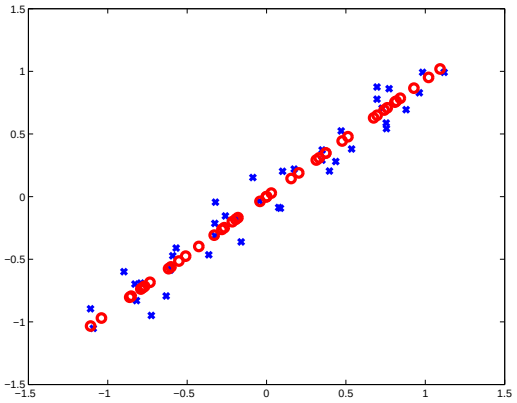

Figure 23.1 A set of vectors in $\mathbb{R}^2$ (blue x's) and their reconstruction after dimensionality reduction to $\mathbb{R}^1$ using PCA (red circles).

PCA

input
A matrix of m examples $X \in \mathbb{R}^{m,d}$
number of components n
if $(m > d)$
$A = X^\top X$
Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be the eigenvectors of A with largest eigenvalues
else
$B = XX^\top$
Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be the eigenvectors of B with largest eigenvalues
for $i = 1, \ldots, n$ set $\mathbf{u}_i = \frac{1}{\|X^\top \mathbf{v}_i\|} X^\top \mathbf{v}_i$
output: $\mathbf{u}_1, \ldots, \mathbf{u}_n$

To illustrate how PCA works, let us generate vectors in $\mathbb{R}^2$ that approximately reside on a line, namely, on a one dimensional subspace of $\mathbb{R}^2$. For example, suppose that each example is of the form $(x, x + y)$ where x is chosen uniformly at random from $[-1, 1]$ and y is sampled from a Gaussian distribution with mean 0 and standard deviation of 0.1. Suppose we apply PCA on this data. Then, the eigenvector corresponding to the largest eigenvalue will be close to the vector $(1/\sqrt{2}, 1/\sqrt{2})$. When projecting a point $(x, x + y)$ on this principal component we will obtain the scalar $\frac{2x+y}{\sqrt{2}}$. The reconstruction of the original vector will be $((x + y/2), (x + y/2))$. In Figure 23.1 we depict the original versus reconstructed data.

Next, we demonstrate the effectiveness of PCA on a data set of faces. We extracted images of faces from the Yale data set (Georghiades, Belhumeur & Kriegman 2001). Each image contains $50 \times 50 = 2500$ pixels; therefore the original dimensionality is very high.

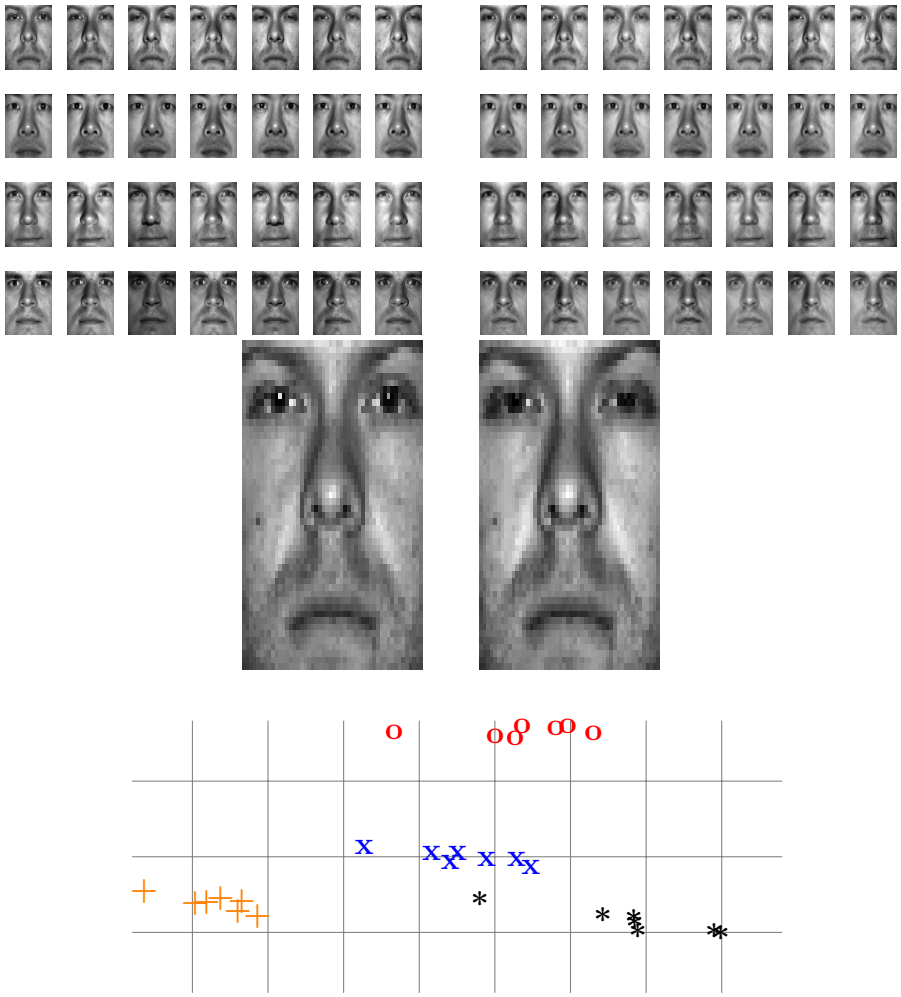

Figure 23.2 Images of faces extracted from the Yale data set. Top-Left: the original images in $\mathbb{R}^{50x50}$. Top-Right: the images after dimensionality reduction to $\mathbb{R}^{10}$ and reconstruction. Middle row: an enlarged version of one of the images before and after PCA. Bottom: The images after dimensionality reduction to $\mathbb{R}^2$. The different marks indicate different individuals.

Some images of faces are depicted on the top-left side of Figure 23.2. Using PCA, we reduced the dimensionality to $\mathbb{R}^{10}$ and reconstructed back to the original dimension, which is 50^2. The resulting reconstructed images are depicted on the top-right side of Figure 23.2. Finally, on the bottom of Figure 23.2 we depict a 2 dimensional representation of the images. As can be seen, even from a 2 dimensional representation of the images we can still roughly separate different individuals.

23.2 Random Projections

In this section we show that reducing the dimension by using a random linear transformation leads to a simple compression scheme with a surprisingly low distortion. The transformation $\mathbf{x} \mapsto W\mathbf{x}$, when W is a random matrix, is often referred to as a random projection. In particular, we provide a variant of a famous lemma due to Johnson and Lindenstrauss, showing that random projections do not distort Euclidean distances too much.

Let $\mathbf{x}_1, \mathbf{x}_2$ be two vectors in $\mathbb{R}^d$. A matrix W does not distort too much the distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ if the ratio

$$\frac{\|W\mathbf{x}_1 - W\mathbf{x}_2\|}{\|\mathbf{x}_1 - \mathbf{x}_2\|}$$

is close to 1. In other words, the distances between $\mathbf{x}_1$ and $\mathbf{x}_2$ before and after the transformation are almost the same. To show that $\|W\mathbf{x}_1 - W\mathbf{x}_2\|$ is not too far away from $\|\mathbf{x}_1 - \mathbf{x}_2\|$ it suffices to show that W does not distort the norm of the difference vector $\mathbf{x} = \mathbf{x}_1 - \mathbf{x}_2$. Therefore, from now on we focus on the ratio $\frac{\|W\mathbf{x}\|}{\|\mathbf{x}\|}$.

We start with analyzing the distortion caused by applying a random projection to a single vector.

LEMMA 23.3 *Fix some $\mathbf{x} \in \mathbb{R}^d$. Let $W \in \mathbb{R}^{n,d}$ be a random matrix such that each $W_{i,j}$ is an independent normal random variable. Then, for every $\epsilon \in (0,3)$ we have*

$$\mathbb{P}\left[\left|\frac{\|(1/\sqrt{n})W\mathbf{x}\|^2}{\|\mathbf{x}\|^2} - 1\right| > \epsilon\right] \leq 2\,e^{-\epsilon^2 n/6}.$$

Proof Without loss of generality we can assume that $\|\mathbf{x}\|^2 = 1$. Therefore, an equivalent inequality is

$$\mathbb{P}\left[(1-\epsilon)n \leq \|W\mathbf{x}\|^2 \leq (1+\epsilon)n\right] \geq 1 - 2e^{-\epsilon^2 n/6}.$$

Let $\mathbf{w}_i$ be the ith row of W. The random variable $\langle \mathbf{w}_i, \mathbf{x}\rangle$ is a weighted sum of d independent normal random variables and therefore it is normally distributed with zero mean and variance $\sum_j x_j^2 = \|\mathbf{x}\|^2 = 1$. Therefore, the random variable $\|W\mathbf{x}\|^2 = \sum_{i=1}^n (\langle \mathbf{w}_i, \mathbf{x}\rangle)^2$ has a χ_n^2 distribution. The claim now follows directly from a measure concentration property of χ^2 random variables stated in Lemma B.12 given in Section B.7. □

The Johnson-Lindenstrauss lemma follows from this using a simple union bound argument.

LEMMA 23.4 (Johnson-Lindenstrauss Lemma) *Let Q be a finite set of vectors in $\mathbb{R}^d$. Let $\delta \in (0,1)$ and n be an integer such that*

$$\epsilon = \sqrt{\frac{6\,\log(2|Q|/\delta)}{n}} \leq 3.$$

Then, with probability of at least $1-\delta$ *over a choice of a random matrix* $W \in \mathbb{R}^{n,d}$ *such that each element of* W *is distributed normally with zero mean and variance of* $1/n$ *we have*

$$\sup_{\mathbf{x}\in Q}\left|\frac{\|W\mathbf{x}\|^2}{\|\mathbf{x}\|^2}-1\right|<\epsilon.$$

Proof Combining Lemma 23.3 and the union bound we have that for every $\epsilon \in (0,3)$:

$$\mathbb{P}\left[\sup_{\mathbf{x}\in Q}\left|\frac{\|W\mathbf{x}\|^2}{\|\mathbf{x}\|^2}-1\right|>\epsilon\right] \leq 2\,|Q|\,e^{-\epsilon^2 n/6}.$$

Let δ denote the right-hand side of the inequality; thus we obtain that

$$\epsilon=\sqrt{\frac{6\,\log(2|Q|/\delta)}{n}}.$$

□

Interestingly, the bound given in Lemma 23.4 does not depend on the original dimension of $\mathbf{x}$. In fact, the bound holds even if $\mathbf{x}$ is in an infinite dimensional Hilbert space.

23.3 Compressed Sensing

Compressed sensing is a dimensionality reduction technique which utilizes a prior assumption that the original vector is sparse in some basis. To motivate compressed sensing, consider a vector $\mathbf{x} \in \mathbb{R}^d$ that has at most s nonzero elements. That is,

$$\|\mathbf{x}\|_0 \stackrel{\text{def}}{=} |\{i : x_i \neq 0\}| \leq s.$$

Clearly, we can compress $\mathbf{x}$ by representing it using s (index,value) pairs. Furthermore, this compression is lossless – we can reconstruct $\mathbf{x}$ exactly from the s (index,value) pairs. Now, lets take one step forward and assume that $\mathbf{x} = U\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is a sparse vector, $\|\boldsymbol{\alpha}\|_0 \leq s$, and U is a fixed orthonormal matrix. That is, $\mathbf{x}$ has a sparse representation in another basis. It turns out that many natural vectors are (at least approximately) sparse in some representation. In fact, this assumption underlies many modern compression schemes. For example, the JPEG-2000 format for image compression relies on the fact that natural images are approximately sparse in a wavelet basis.

Can we still compress $\mathbf{x}$ into roughly s numbers? Well, one simple way to do this is to multiply $\mathbf{x}$ by $U^\top$, which yields the sparse vector $\boldsymbol{\alpha}$, and then represent $\boldsymbol{\alpha}$ by its s (index,value) pairs. However, this requires us first to "sense" $\mathbf{x}$, to store it, and then to multiply it by $U^\top$. This raises a very natural question: Why go to so much effort to acquire all the data when most of what we get will be thrown away? Cannot we just directly measure the part that will not end up being thrown away?

Compressed sensing is a technique that simultaneously acquires and compresses the data. The key result is that a random linear transformation can compress $\mathbf{x}$ without losing information. The number of measurements needed is order of $s \log(d)$. That is, we roughly acquire only the important information about the signal. As we will see later, the price we pay is a slower reconstruction phase. In some situations, it makes sense to save time in compression even at the price of a slower reconstruction. For example, a security camera should sense and compress a large amount of images while most of the time we do not need to decode the compressed data at all. Furthermore, in many practical applications, compression by a linear transformation is advantageous because it can be performed efficiently in hardware. For example, a team led by Baraniuk and Kelly has proposed a camera architecture that employs a digital micromirror array to perform optical calculations of a linear transformation of an image. In this case, obtaining each compressed measurement is as easy as obtaining a single raw measurement. Another important application of compressed sensing is medical imaging, in which requiring fewer measurements translates to less radiation for the patient.

Informally, the main premise of compressed sensing is the following three "surprising" results:

1. It is possible to reconstruct any sparse signal fully if it was compressed by $\mathbf{x} \mapsto W\mathbf{x}$, where W is a matrix which satisfies a condition called the Restricted Isoperimetric Property (RIP). A matrix that satisfies this property is guaranteed to have a low distortion of the norm of any sparse representable vector.
2. The reconstruction can be calculated in polynomial time by solving a linear program.
3. A random $n \times d$ matrix is likely to satisfy the RIP condition provided that n is greater than an order of $s \log(d)$.

Formally,

DEFINITION 23.5 (RIP) A matrix $W \in \mathbb{R}^{n,d}$ is (ϵ, s)-RIP if for all $\mathbf{x} \neq 0$ s.t. $\|\mathbf{x}\|_0 \leq s$ we have

$$\left| \frac{\|W\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} - 1 \right| \leq \epsilon.$$

The first theorem establishes that RIP matrices yield a lossless compression scheme for sparse vectors. It also provides a (nonefficient) reconstruction scheme.

THEOREM 23.6 *Let $\epsilon < 1$ and let W be a $(\epsilon, 2s)$-RIP matrix. Let $\mathbf{x}$ be a vector s.t. $\|\mathbf{x}\|_0 \leq s$, let $\mathbf{y} = W\mathbf{x}$ be the compression of $\mathbf{x}$, and let*

$$\tilde{\mathbf{x}} \in \operatorname*{argmin}_{\mathbf{v}: W\mathbf{v}=\mathbf{y}} \|\mathbf{v}\|_0$$

be a reconstructed vector. Then, $\tilde{\mathbf{x}} = \mathbf{x}$.

Proof We assume, by way of contradiction, that $\tilde{\mathbf{x}} \neq \mathbf{x}$. Since $\mathbf{x}$ satisfies the constraints in the optimization problem that defines $\tilde{\mathbf{x}}$ we clearly have that $\|\tilde{\mathbf{x}}\|_0 \leq \|\mathbf{x}\|_0 \leq s$. Therefore, $\|\mathbf{x} - \tilde{\mathbf{x}}\|_0 \leq 2s$ and we can apply the RIP inequality on the vector $\mathbf{x} - \tilde{\mathbf{x}}$. But, since $W(\mathbf{x} - \tilde{\mathbf{x}}) = \mathbf{0}$ we get that $|0 - 1| \leq \epsilon$, which leads to a contradiction. □

The reconstruction scheme given in Theorem 23.6 seems to be nonefficient because we need to minimize a combinatorial objective (the sparsity of $\mathbf{v}$). Quite surprisingly, it turns out that we can replace the combinatorial objective, $\|\mathbf{v}\|_0$, with a convex objective, $\|\mathbf{v}\|_1$, which leads to a linear programming problem that can be solved efficiently. This is stated formally in the following theorem.

THEOREM 23.7 *Assume that the conditions of Theorem 23.6 holds and that* $\epsilon < \frac{1}{1+\sqrt{2}}$. *Then,*

$$\mathbf{x} = \operatorname*{argmin}_{\mathbf{v}: W\mathbf{v}=\mathbf{y}} \|\mathbf{v}\|_0 = \operatorname*{argmin}_{\mathbf{v}: W\mathbf{v}=\mathbf{y}} \|\mathbf{v}\|_1.$$

In fact, we will prove a stronger result, which holds even if $\mathbf{x}$ is not a sparse vector.

THEOREM 23.8 *Let* $\epsilon < \frac{1}{1+\sqrt{2}}$ *and let* W *be a* $(\epsilon, 2s)$*-RIP matrix. Let* $\mathbf{x}$ *be an arbitrary vector and denote*

$$\mathbf{x}_s \in \operatorname*{argmin}_{\mathbf{v}: \|\mathbf{v}\|_0 \leq s} \|\mathbf{x} - \mathbf{v}\|_1.$$

That is, $\mathbf{x}_s$ *is the vector which equals* $\mathbf{x}$ *on the* s *largest elements of* $\mathbf{x}$ *and equals* 0 *elsewhere. Let* $\mathbf{y} = W\mathbf{x}$ *be the compression of* $\mathbf{x}$ *and let*

$$\mathbf{x}^\star \in \operatorname*{argmin}_{\mathbf{v}: W\mathbf{v}=\mathbf{y}} \|\mathbf{v}\|_1$$

be the reconstructed vector. Then,

$$\|\mathbf{x}^\star - \mathbf{x}\|_2 \leq 2\frac{1+\rho}{1-\rho} s^{-1/2} \|\mathbf{x} - \mathbf{x}_s\|_1,$$

where $\rho = \sqrt{2}\epsilon/(1-\epsilon)$.

Note that in the special case that $\mathbf{x} = \mathbf{x}_s$ we get an exact recovery, $\mathbf{x}^\star = \mathbf{x}$, so Theorem 23.7 is a special case of Theorem 23.8. The proof of Theorem 23.8 is given in Section 23.3.1.

Finally, the third result tells us that random matrices with $n \geq \Omega(s \log(d))$ are likely to be RIP. In fact, the theorem shows that multiplying a random matrix by an orthonormal matrix also provides an RIP matrix. This is important for compressing signals of the form $\mathbf{x} = U\boldsymbol{\alpha}$ where $\mathbf{x}$ is not sparse but $\boldsymbol{\alpha}$ is sparse. In that case, if W is a random matrix and we compress using $\mathbf{y} = W\mathbf{x}$ then this is the same as compressing $\boldsymbol{\alpha}$ by $\mathbf{y} = (WU)\boldsymbol{\alpha}$ and since WU is also RIP we can reconstruct $\boldsymbol{\alpha}$ (and thus also $\mathbf{x}$) from $\mathbf{y}$.

THEOREM 23.9 *Let U be an arbitrary fixed $d \times d$ orthonormal matrix, let ϵ, δ be scalars in $(0,1)$, let s be an integer in $[d]$, and let n be an integer that satisfies*

$$n \geq 100 \frac{s \log(40d/(\delta\,\epsilon))}{\epsilon^2}.$$

Let $W \in \mathbb{R}^{n,d}$ be a matrix s.t. each element of W is distributed normally with zero mean and variance of $1/n$. Then, with proabability of at least $1-\delta$ over the choice of W, the matrix WU is (ϵ, s)-RIP.

23.3.1 Proofs*

Proof of Theorem 23.8

We follow a proof due to Candès (2008).

Let $\mathbf{h} = \mathbf{x}^\star - \mathbf{x}$. Given a vector $\mathbf{v}$ and a set of indices I we denote by $\mathbf{v}_I$ the vector whose ith element is v_i if $i \in I$ and 0 otherwise.

The first trick we use is to partition the set of indices $[d] = \{1, \ldots, d\}$ into disjoint sets of size s. That is, we will write $[d] = T_0 \cup T_1 \cup T_2 \ldots T_{d/s-1}$ where for all i, $|T_i| = s$, and we assume for simplicity that d/s is an integer. We define the partition as follows. In T_0 we put the s indices corresponding to the s largest elements in absolute values of $\mathbf{x}$ (ties are broken arbitrarily). Let $T_0^c = [d] \setminus T_0$. Next, T_1 will be the s indices corresponding to the s largest elements in absolute value of $\mathbf{h}_{T_0^c}$. Let $T_{0,1} = T_0 \cup T_1$ and $T_{0,1}^c = [d] \setminus T_{0,1}$. Next, T_2 will correspond to the s largest elements in absolute value of $\mathbf{h}_{T_{0,1}^c}$. And, we will construct $T_3, T_4, \ldots$ in the same way.

To prove the theorem we first need the following lemma, which shows that RIP also implies approximate orthogonality.

LEMMA 23.10 *Let W be an $(\epsilon, 2s)$-RIP matrix. Then, for any two disjoint sets I, J, both of size at most s, and for any vector $\mathbf{u}$ we have that $\langle W\mathbf{u}_I, W\mathbf{u}_J \rangle \leq \epsilon \|\mathbf{u}_I\|_2 \, \|\mathbf{u}_J\|_2$.*

Proof W.l.o.g. assume $\|\mathbf{u}_I\|_2 = \|\mathbf{u}_J\|_2 = 1$.

$$\langle W\mathbf{u}_I, W\mathbf{u}_J \rangle = \frac{\|W\mathbf{u}_I + W\mathbf{u}_J\|_2^2 - \|W\mathbf{u}_I - W\mathbf{u}_J\|_2^2}{4}.$$

But, since $|J \cup I| \leq 2s$ we get from the RIP condition that $\|W\mathbf{u}_I + W\mathbf{u}_J\|_2^2 \leq (1+\epsilon)(\|\mathbf{u}_I\|_2^2 + \|\mathbf{u}_J\|_2^2) = 2(1+\epsilon)$ and that $-\|W\mathbf{u}_I - W\mathbf{u}_J\|_2^2 \leq -(1-\epsilon)(\|\mathbf{u}_I\|_2^2 + \|\mathbf{u}_J\|_2^2) = -2(1-\epsilon)$, which concludes our proof. □

We are now ready to prove the theorem. Clearly,

$$\|\mathbf{h}\|_2 = \|\mathbf{h}_{T_{0,1}} + \mathbf{h}_{T_{0,1}^c}\|_2 \leq \|\mathbf{h}_{T_{0,1}}\|_2 + \|\mathbf{h}_{T_{0,1}^c}\|_2. \tag{23.5}$$

To prove the theorem we will show the following two claims:

Claim 1:. $\|\mathbf{h}_{T_{0,1}^c}\|_2 \leq \|\mathbf{h}_{T_0}\|_2 + 2s^{-1/2}\|\mathbf{x} - \mathbf{x}_s\|_1$.

Claim 2:. $\|\mathbf{h}_{T_{0,1}}\|_2 \leq \frac{2\rho}{1-\rho} s^{-1/2}\|\mathbf{x} - \mathbf{x}_s\|_1$.

Combining these two claims with Equation (23.5) we get that

$$\begin{aligned}\|\mathbf{h}\|_2 &\leq \|\mathbf{h}_{T_{0,1}}\|_2 + \|\mathbf{h}_{T_{0,1}^c}\|_2 \leq 2\|\mathbf{h}_{T_{0,1}}\|_2 + 2s^{-1/2}\|\mathbf{x}-\mathbf{x}_s\|_1 \\ &\leq 2\left(\tfrac{2\rho}{1-\rho}+1\right)s^{-1/2}\|\mathbf{x}-\mathbf{x}_s\|_1 \\ &= 2\frac{1+\rho}{1-\rho}\,s^{-1/2}\|\mathbf{x}-\mathbf{x}_s\|_1,\end{aligned}$$

and this will conclude our proof.

Proving Claim 1:
To prove this claim we do not use the RIP condition at all but only use the fact that $\mathbf{x}^\star$ minimizes the ℓ_1 norm. Take $j > 1$. For each $i \in T_j$ and $i' \in T_{j-1}$ we have that $|h_i| \leq |h_{i'}|$. Therefore, $\|\mathbf{h}_{T_j}\|_\infty \leq \|\mathbf{h}_{T_{j-1}}\|_1/s$. Thus,

$$\|\mathbf{h}_{T_j}\|_2 \leq s^{1/2}\|\mathbf{h}_{T_j}\|_\infty \leq s^{-1/2}\|\mathbf{h}_{T_{j-1}}\|_1.$$

Summing this over $j = 2, 3, \ldots$ and using the triangle inequality we obtain that

$$\|\mathbf{h}_{T_{0,1}^c}\|_2 \leq \sum_{j\geq 2}\|\mathbf{h}_{T_j}\|_2 \leq s^{-1/2}\|\mathbf{h}_{T_0^c}\|_1 \tag{23.6}$$

Next, we show that $\|\mathbf{h}_{T_0^c}\|_1$ cannot be large. Indeed, from the definition of $\mathbf{x}^\star$ we have that $\|\mathbf{x}\|_1 \geq \|\mathbf{x}^\star\|_1 = \|\mathbf{x}+\mathbf{h}\|_1$. Thus, using the triangle inequality we obtain that

$$\|\mathbf{x}\|_1 \geq \|\mathbf{x}+\mathbf{h}\|_1 = \sum_{i\in T_0}|x_i+h_i| + \sum_{i\in T_0^c}|x_i+h_i| \geq \|\mathbf{x}_{T_0}\|_1 - \|\mathbf{h}_{T_0}\|_1 + \|\mathbf{h}_{T_0^c}\|_1 - \|\mathbf{x}_{T_0^c}\|_1 \tag{23.7}$$

and since $\|\mathbf{x}_{T_0^c}\|_1 = \|\mathbf{x}-\mathbf{x}_s\|_1 = \|\mathbf{x}\|_1 - \|\mathbf{x}_{T_0}\|_1$ we get that

$$\|\mathbf{h}_{T_0^c}\|_1 \leq \|\mathbf{h}_{T_0}\|_1 + 2\|\mathbf{x}_{T_0^c}\|_1. \tag{23.8}$$

Combining this with Equation (23.6) we get that

$$\|\mathbf{h}_{T_{0,1}^c}\|_2 \leq s^{-1/2}\left(\|\mathbf{h}_{T_0}\|_1 + 2\|\mathbf{x}_{T_0^c}\|_1\right) \leq \|\mathbf{h}_{T_0}\|_2 + 2s^{-1/2}\|\mathbf{x}_{T_0^c}\|_1,$$

which concludes the proof of claim 1.

Proving Claim 2:
For the second claim we use the RIP condition to get that

$$(1-\epsilon)\|\mathbf{h}_{T_{0,1}}\|_2^2 \leq \|W\mathbf{h}_{T_{0,1}}\|_2^2. \tag{23.9}$$

Since $W\mathbf{h}_{T_{0,1}} = W\mathbf{h} - \sum_{j\geq 2}W\mathbf{h}_{T_j} = -\sum_{j\geq 2}W\mathbf{h}_{T_j}$ we have that

$$\|W\mathbf{h}_{T_{0,1}}\|_2^2 = -\sum_{j\geq 2}\langle W\mathbf{h}_{T_{0,1}}, W\mathbf{h}_{T_j}\rangle = -\sum_{j\geq 2}\langle W\mathbf{h}_{T_0} + W\mathbf{h}_{T_1}, W\mathbf{h}_{T_j}\rangle.$$

From the RIP condition on inner products we obtain that for all $i \in \{1, 2\}$ and $j \geq 2$ we have

$$|\langle W\mathbf{h}_{T_i}, W\mathbf{h}_{T_j}\rangle| \leq \epsilon\|\mathbf{h}_{T_i}\|_2\|\mathbf{h}_{T_j}\|_2.$$

Since $\|\mathbf{h}_{T_0}\|_2 + \|\mathbf{h}_{T_1}\|_2 \leq \sqrt{2}\|\mathbf{h}_{T_{0,1}}\|_2$ we therefore get that

$$\|W\mathbf{h}_{T_{0,1}}\|_2^2 \;\leq\; \sqrt{2}\epsilon\|\mathbf{h}_{T_{0,1}}\|_2 \sum_{j\geq 2}\|\mathbf{h}_{T_j}\|_2.$$

Combining this with Equation (23.6) and Equation (23.9) we obtain

$$(1-\epsilon)\|\mathbf{h}_{T_{0,1}}\|_2^2 \leq \sqrt{2}\epsilon\|\mathbf{h}_{T_{0,1}}\|_2 s^{-1/2}\|\mathbf{h}_{T_0^c}\|_1.$$

Rearranging the inequality gives

$$\|\mathbf{h}_{T_{0,1}}\|_2 \leq \frac{\sqrt{2}\epsilon}{1-\epsilon} s^{-1/2}\|\mathbf{h}_{T_0^c}\|_1.$$

Finally, using Equation (23.8) we get that

$$\|\mathbf{h}_{T_{0,1}}\|_2 \leq \rho s^{-1/2}\left(\|\mathbf{h}_{T_0}\|_1 + 2\|\mathbf{x}_{T_0^c}\|_1\right) \leq \rho\|\mathbf{h}_{T_0}\|_2 + 2\rho s^{-1/2}\|\mathbf{x}_{T_0^c}\|_1,$$

but since $\|\mathbf{h}_{T_0}\|_2 \leq \|\mathbf{h}_{T_{0,1}}\|_2$ this implies

$$\|\mathbf{h}_{T_{0,1}}\|_2 \leq \frac{2\rho}{1-\rho}\, s^{-1/2}\|\mathbf{x}_{T_0^c}\|_1,$$

which concludes the proof of the second claim.

Proof of Theorem 23.9

To prove the theorem we follow an approach due to (Baraniuk, Davenport, DeVore & Wakin 2008). The idea is to combine the Johnson-Lindenstrauss (JL) lemma with a simple covering argument.

We start with a covering property of the unit ball.

LEMMA 23.11 *Let $\epsilon \in (0,1)$. There exists a finite set $Q \subset \mathbb{R}^d$ of size $|Q| \leq \left(\frac{3}{\epsilon}\right)^d$ such that*

$$\sup_{\mathbf{x}:\|\mathbf{x}\|\leq 1} \min_{\mathbf{v}\in Q} \|\mathbf{x}-\mathbf{v}\| \;\leq\; \epsilon.$$

Proof Let k be an integer and let

$$Q' = \{\mathbf{x} \in \mathbb{R}^d : \forall j \in [d], \exists i \in \{-k, -k+1, \ldots, k\} \text{ s.t. } x_j = \tfrac{i}{k}\}.$$

Clearly, $|Q'| = (2k+1)^d$. We shall set $Q = Q' \cap B_2(1)$, where $B_2(1)$ is the unit ℓ_2 ball of $\mathbb{R}^d$. Since the points in Q' are distributed evenly on the unit ℓ_∞ ball, the size of Q is the size of Q' times the ratio between the volumes of the unit ℓ_2 and ℓ_∞ balls. The volume of the ℓ_∞ ball is 2^d and the volume of $B_2(1)$ is

$$\frac{\pi^{d/2}}{\Gamma(1+d/2)}.$$

For simplicity, assume that d is even and therefore

$$\Gamma(1+d/2) \;=\; (d/2)! \;\geq\; \left(\tfrac{d/2}{e}\right)^{d/2},$$

where in the last inequality we used Stirling's approximation. Overall we obtained that

$$|Q| \leq (2k+1)^d \, (\pi/e)^{d/2} \, (d/2)^{-d/2} \, 2^{-d}. \tag{23.10}$$

Now lets specify k. For each $\mathbf{x} \in B_2(1)$ let $\mathbf{v} \in Q$ be the vector whose ith element is $\text{sign}(x_i) \, \lfloor |x_i| \, k \rfloor / k$. Then, for each element we have that $|x_i - v_i| \leq 1/k$ and thus

$$\|\mathbf{x} - \mathbf{v}\| \leq \frac{\sqrt{d}}{k}.$$

To ensure that the right-hand side will be at most ϵ we shall set $k = \lceil \sqrt{d}/\epsilon \rceil$. Plugging this value into Equation (23.10) we conclude that

$$|Q| \leq (3\sqrt{d}/(2\epsilon))^d \, (\pi/e)^{d/2} \, (d/2)^{-d/2} = \left(\tfrac{3}{\epsilon} \sqrt{\tfrac{\pi}{2e}} \right)^d \leq \left(\tfrac{3}{\epsilon} \right)^d.$$

□

Let $\mathbf{x}$ be a vector that can be written as $\mathbf{x} = U\boldsymbol{\alpha}$ with U being some orthonormal matrix and $\|\boldsymbol{\alpha}\|_0 \leq s$. Combining the earlier covering property and the JL lemma (Lemma 23.4) enables us to show that a random W will not distort any such $\mathbf{x}$.

LEMMA 23.12 *Let U be an orthonormal $d \times d$ matrix and let $I \subset [d]$ be a set of indices of size $|I| = s$. Let S be the span of $\{U_i : i \in I\}$, where U_i is the ith column of U. Let $\delta \in (0,1)$, $\epsilon \in (0,1)$, and $n \in \mathbb{N}$ such that*

$$n \geq 24 \, \frac{\log(2/\delta) + s \, \log(12/\epsilon)}{\epsilon^2}.$$

Then, with probability of at least $1-\delta$ over a choice of a random matrix $W \in \mathbb{R}^{n,d}$ such that each element of W is independently distributed according to $N(0, 1/n)$, we have

$$\sup_{\mathbf{x} \in S} \left| \frac{\|W\mathbf{x}\|}{\|\mathbf{x}\|} - 1 \right| < \epsilon.$$

Proof It suffices to prove the lemma for all $\mathbf{x} \in S$ with $\|\mathbf{x}\| = 1$. We can write $\mathbf{x} = U_I \boldsymbol{\alpha}$ where $\boldsymbol{\alpha} \in \mathbb{R}^s$, $\|\boldsymbol{\alpha}\|_2 = 1$, and U_I is the matrix whose columns are $\{U_i : i \in I\}$. Using Lemma 23.11 we know that there exists a set Q of size $|Q| \leq (12/\epsilon)^s$ such that

$$\sup_{\boldsymbol{\alpha}: \|\boldsymbol{\alpha}\|=1} \min_{\mathbf{v} \in Q} \|\boldsymbol{\alpha} - \mathbf{v}\| \leq (\epsilon/4).$$

But since U is orthogonal we also have that

$$\sup_{\boldsymbol{\alpha}: \|\boldsymbol{\alpha}\|=1} \min_{\mathbf{v} \in Q} \|U_I\boldsymbol{\alpha} - U_I\mathbf{v}\| \leq (\epsilon/4).$$

Applying Lemma 23.4 on the set $\{U_I\mathbf{v} : \mathbf{v} \in Q\}$ we obtain that for n satisfying

the condition given in the lemma, the following holds with probability of at least $1-\delta$:

$$\sup_{\mathbf{v}\in Q}\left|\frac{\|WU_I\mathbf{v}\|^2}{\|U_I\mathbf{v}\|^2}-1\right|\leq \epsilon/2,$$

This also implies that

$$\sup_{\mathbf{v}\in Q}\left|\frac{\|WU_I\mathbf{v}\|}{\|U_I\mathbf{v}\|}-1\right|\leq \epsilon/2.$$

Let a be the smallest number such that

$$\forall \mathbf{x}\in S,\quad \frac{\|W\mathbf{x}\|}{\|\mathbf{x}\|}\leq 1+a.$$

Clearly $a<\infty$. Our goal is to show that $a\leq\epsilon$. This follows from the fact that for any $\mathbf{x}\in S$ of unit norm there exists $\mathbf{v}\in Q$ such that $\|\mathbf{x}-U_I\mathbf{v}\|\leq\epsilon/4$ and therefore

$$\|W\mathbf{x}\|\leq\|WU_I\mathbf{v}\|+\|W(\mathbf{x}-U_I\mathbf{v})\|\leq 1+\epsilon/2+(1+a)\epsilon/4.$$

Thus,

$$\forall \mathbf{x}\in S,\quad \frac{\|W\mathbf{x}\|}{\|\mathbf{x}\|}\leq 1+(\epsilon/2+(1+a)\epsilon/4)\,.$$

But the definition of a implies that

$$a\leq\epsilon/2+(1+a)\epsilon/4\quad\Rightarrow\quad a\leq\frac{\epsilon/2+\epsilon/4}{1-\epsilon/4}\leq\epsilon.$$

This proves that for all $\mathbf{x}\in S$ we have $\frac{\|W\mathbf{x}\|}{\|\mathbf{x}\|}-1\leq\epsilon$. The other side follows from this as well since

$$\|W\mathbf{x}\|\geq\|WU_I\mathbf{v}\|-\|W(\mathbf{x}-U_I\mathbf{v})\|\geq 1-\epsilon/2-(1+\epsilon)\epsilon/4\geq 1-\epsilon.$$

□

The preceding lemma tells us that for $\mathbf{x}\in S$ of unit norm we have

$$(1-\epsilon)\leq\|W\mathbf{x}\|\leq(1+\epsilon),$$

which implies that

$$(1-2\,\epsilon)\leq\|W\mathbf{x}\|^2\leq(1+3\,\epsilon).$$

The proof of Theorem 23.9 follows from this by a union bound over all choices of I.

23.4 PCA or Compressed Sensing?

Suppose we would like to apply a dimensionality reduction technique to a given set of examples. Which method should we use, PCA or compressed sensing? In this section we tackle this question, by underscoring the underlying assumptions behind the two methods.

It is helpful first to understand when each of the methods can guarantee perfect recovery. PCA guarantees perfect recovery whenever the set of examples is contained in an n dimensional subspace of $\mathbb{R}^d$. Compressed sensing guarantees perfect recovery whenever the set of examples is sparse (in some basis). On the basis of these observations, we can describe cases in which PCA will be better than compressed sensing and vice versa.

As a first example, suppose that the examples are the vectors of the standard basis of $\mathbb{R}^d$, namely, $\mathbf{e}_1, \ldots, \mathbf{e}_d$, where each $\mathbf{e}_i$ is the all zeros vector except 1 in the ith coordinate. In this case, the examples are 1-sparse. Hence, compressed sensing will yield a perfect recovery whenever $n \geq \Omega(\log(d))$. On the other hand, PCA will lead to poor performance, since the data is far from being in an n dimensional subspace, as long as $n < d$. Indeed, it is easy ro verify that in such a case, the averaged recovery error of PCA (i.e., the objective of Equation (23.1) divided by m) will be $(d-n)/d$, which is larger than $1/2$ whenever $n \leq d/2$.

We next show a case where PCA is better than compressed sensing. Consider m examples that are exactly on an n dimensional subspace. Clearly, in such a case, PCA will lead to perfect recovery. As to compressed sensing, note that the examples are n-sparse in any orthonormal basis whose first n vectors span the subspace. Therefore, compressed sensing would also work if we will reduce the dimension to $\Omega(n \log(d))$. However, with exactly n dimensions, compressed sensing might fail. PCA has also better resilience to certain types of noise. See (Chang, Weiss & Freeman 2009) for a discussion.

23.5 Summary

We introduced two methods for dimensionality reduction using linear transformations: PCA and random projections. We have shown that PCA is optimal in the sense of averaged squared reconstruction error, if we restrict the reconstruction procedure to be linear as well. However, if we allow nonlinear reconstruction, PCA is not necessarily the optimal procedure. In particular, for sparse data, random projections can significantly outperform PCA. This fact is at the heart of the compressed sensing method.

23.6 Bibliographic Remarks

PCA is equivalent to best subspace approximation using singular value decomposition (SVD). The SVD method is described in Appendix C. SVD dates back to Eugenio Beltrami (1873) and Camille Jordan (1874). It has been rediscovered many times. In the statistical literature, it was introduced by Pearson (1901). Besides PCA and SVD, there are additional names that refer to the same idea and are being used in different scientific communities. A few examples are the Eckart-Young theorem (after Carl Eckart and Gale Young who analyzed the method in 1936), the Schmidt-Mirsky theorem, factor analysis, and the Hotelling transform.

Compressed sensing was introduced in Donoho (2006) and in (Candes & Tao 2005). See also Candes (2006).

23.7 Exercises

1. In this exercise we show that in the general case, exact recovery of a linear compression scheme is impossible.
   1. let $A \in \mathbb{R}^{n,d}$ be an arbitrary compression matrix where $n \leq d-1$. Show that there exists $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, $\mathbf{u} \neq \mathbf{v}$ such that $A\mathbf{u} = A\mathbf{v}$.
   2. Conclude that exact recovery of a linear compression scheme is impossible.
2. Let $\boldsymbol{\alpha} \in \mathbb{R}^d$ such that $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_d \geq 0$. Show that

$$\max_{\boldsymbol{\beta} \in [0,1]^d : \|\boldsymbol{\beta}\|_1 \leq n} \sum_{j=1}^{d} \alpha_j \beta_j \quad = \quad \sum_{j=1}^{n} \alpha_j .$$

   Hint: Take every vector $\boldsymbol{\beta} \in [0,1]^d$ such that $\|\boldsymbol{\beta}\|_1 \leq n$. Let i be the minimal index for which $\beta_i < 1$. If $i = n+1$ we are done. Otherwise, show that we can increase β_i, while possibly decreasing β_j for some $j > i$, and obtain a better solution. This will imply that the optimal solution is to set $\beta_i = 1$ for $i \leq n$ and $\beta_i = 0$ for $i > n$.
3. **Kernel PCA:** In this exercise we show how PCA can be used for constructing nonlinear dimensionality reduction on the basis of the kernel trick (see Chapter 16).

   Let $\mathcal{X}$ be some instance space and let $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be a set of points in $\mathcal{X}$. Consider a feature mapping $\psi : \mathcal{X} \to V$, where V is some Hilbert space (possibly of infinite dimension). Let $K : \mathcal{X} \times \mathcal{X}$ be a kernel function, that is, $k(\mathbf{x}, \mathbf{x}') = \langle \psi(\mathbf{x}), \psi(\mathbf{x}') \rangle$. Kernel PCA is the process of mapping the elements in S into V using ψ, and then applying PCA over $\{\psi(\mathbf{x}_1), \ldots, \psi(\mathbf{x}_m)\}$ into $\mathbb{R}^n$. The output of this process is the set of reduced elements.

   Show how this process can be done in polynomial time in terms of m and n, assuming that each evaluation of $K(\cdot, \cdot)$ can be calculated in a constant time. In particular, if your implementation requires multiplication of two matrices A and B, verify that their product can be computed. Similarly,

if an eigenvalue decomposition of some matrix C is required, verify that this decomposition can be computed.

4. **An Interpretation of PCA as Variance Maximization:**

   Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be m vectors in $\mathbb{R}^d$, and let $\mathbf{x}$ be a random vector distributed according to the uniform distribution over $\mathbf{x}_1, \ldots, \mathbf{x}_m$. Assume that $\mathbb{E}[\mathbf{x}] = \mathbf{0}$.

   1. Consider the problem of finding a unit vector, $\mathbf{w} \in \mathbb{R}^d$, such that the random variable $\langle \mathbf{w}, \mathbf{x} \rangle$ has maximal variance. That is, we would like to solve the problem

      $$\operatorname*{argmax}_{\mathbf{w}:\|\mathbf{w}\|=1} \operatorname{Var}[\langle \mathbf{w}, \mathbf{x} \rangle] = \operatorname*{argmax}_{\mathbf{w}:\|\mathbf{w}\|=1} \frac{1}{m} \sum_{i=1}^{m} (\langle \mathbf{w}, \mathbf{x}_i \rangle)^2.$$

      Show that the solution of the problem is to set $\mathbf{w}$ to be the first principle vector of $\mathbf{x}_1, \ldots, \mathbf{x}_m$.

   2. Let $\mathbf{w}_1$ be the first principal component as in the previous question. Now, suppose we would like to find a second unit vector, $\mathbf{w}_2 \in \mathbb{R}^d$, that maximizes the variance of $\langle \mathbf{w}_2, \mathbf{x} \rangle$, but is also uncorrelated to $\langle \mathbf{w}_1, \mathbf{x} \rangle$. That is, we would like to solve:

      $$\operatorname*{argmax}_{\mathbf{w}:\|\mathbf{w}\|=1,\ \mathbb{E}[(\langle \mathbf{w}_1, \mathbf{x} \rangle)(\langle \mathbf{w}, \mathbf{x} \rangle)]=0} \operatorname{Var}[\langle \mathbf{w}, \mathbf{x} \rangle].$$

      Show that the solution to this problem is to set $\mathbf{w}$ to be the second principal component of $\mathbf{x}_1, \ldots, \mathbf{x}_m$.

      Hint: Note that

      $$\mathbb{E}[(\langle \mathbf{w}_1, \mathbf{x} \rangle)(\langle \mathbf{w}, \mathbf{x} \rangle)] = \mathbf{w}_1^\top \mathbb{E}[\mathbf{x}\mathbf{x}^\top]\mathbf{w} = m\mathbf{w}_1^\top A \mathbf{w},$$

      where $A = \sum_i \mathbf{x}_i \mathbf{x}_i^\top$. Since $\mathbf{w}$ is an eigenvector of A we have that the constraint $\mathbb{E}[(\langle \mathbf{w}_1, \mathbf{x} \rangle)(\langle \mathbf{w}, \mathbf{x} \rangle)] = 0$ is equivalent to the constraint

      $$\langle \mathbf{w}_1, \mathbf{w} \rangle = 0.$$

5. **The Relation between SVD and PCA:** Use the SVD theorem (Corollary C.6) for providing an alternative proof of Theorem 23.2.

6. **Random Projections Preserve Inner Products:** The Johnson-Lindenstrauss lemma tells us that a random projection preserves distances between a finite set of vectors. In this exercise you need to prove that if the set of vectors are within the unit ball, then not only are the distances between any two vectors preserved, but the inner product is also preserved.

   Let Q be a finite set of vectors in $\mathbb{R}^d$ and assume that for every $\mathbf{x} \in Q$ we have $\|\mathbf{x}\| \leq 1$.

   1. Let $\delta \in (0, 1)$ and n be an integer such that

      $$\epsilon = \sqrt{\frac{6 \log(|Q|^2/\delta)}{n}} \leq 3.$$

      Prove that with probability of at least $1 - \delta$ over a choice of a random

matrix $W \in \mathbb{R}^{n,d}$, where each element of W is independently distributed according to $\mathcal{N}(0, 1/n)$, we have

$$|\langle W\mathbf{u}, W\mathbf{v}\rangle - \langle \mathbf{u}, \mathbf{v}\rangle| \leq \epsilon$$

for every $\mathbf{u}, \mathbf{v} \in Q$.
Hint: Use JL to bound both $\frac{\|W(\mathbf{u}+\mathbf{v})\|}{\|\mathbf{u}+\mathbf{v}\|}$ and $\frac{\|W(\mathbf{u}-\mathbf{v})\|}{\|\mathbf{u}-\mathbf{v}\|}$.

2. **(*)** Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be a set of vectors in $\mathbb{R}^d$ of norm at most 1, and assume that these vectors are linearly separable with margin of γ. Assume that $d \gg 1/\gamma^2$. Show that there exists a constant $c > 0$ such that if we randomly project these vectors into $\mathbb{R}^n$, for $n = c/\gamma^2$, then with probability of at least 99% it holds that the projected vectors are linearly separable with margin $\gamma/2$.