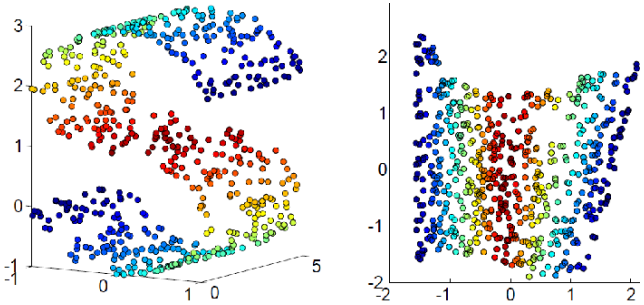
Lecture 10. Dimension Reduction

Bao Wang
Department of Mathematics
Scientific Computing and Imaging Institute
University of Utah
Math 5750/6880, Fall 2023

• **Dimension reduction:** mapping data in a high dimensional space into a new space whose dimensionality is much smaller.



Linear dimension reduction: 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 $\boldsymbol{W} \in \mathbb{R}^{n \times d}$ that induces the mapping $\boldsymbol{x} \rightarrow \boldsymbol{W}\boldsymbol{x}$.

Outline

- Random projection
- Principal component analysis (PCA)
- Compressed sensing

- Random projection: probably approximately preserves the pairwise distance

- PCA: the high-dimensional data can be nearly recovered from the reduced data

- Compressed sensing: acquire reduced data and reconstruct the complete data from the reduced data

Random Projections

- The first natural criterion for choosing $\boldsymbol{W}$ is in a way such that the pairwise distance between reduced data is probably approximately the same as that of the original data.

$$\mathbb{P}\left[\left|\frac{\|\boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{x}_j\|^2}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2} - 1\right| > \epsilon\right] \leq o(\epsilon)$$

• We show that reducing the dimension by using a random linear transformation leads to a simple compression scheme with surprisingly low distortion. — Nearly preserving pairwise distance.

• The transformation $\boldsymbol{x} \rightarrow \boldsymbol{W}\boldsymbol{x}$, when $\boldsymbol{W}$ is a random matrix (each $W_{i,j}$ is an independent normal random variable), is often referred to as a random projection.

• Let $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$. A matrix $\boldsymbol{W} \in \mathbb{R}^{n \times d}$ with $n < d$ does not distort too much the distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ if the ratio

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

is close to 1.

• In other words, the distances between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ before and after the transformation are almost the same. To show that $\|\boldsymbol{W}\boldsymbol{x}_1 - \boldsymbol{W}\boldsymbol{x}_2\|$ is not too far away from $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$ it suffices to show that $\boldsymbol{W}$ does not distort the norm of the difference vector $\boldsymbol{x} = \boldsymbol{x}_1 - \boldsymbol{x}_2$.

• Therefore, from now on we focus on the ratio $\frac{\|\boldsymbol{W}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}$.

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

• **Lemma 2.** Fix some $\boldsymbol{x} \in \mathbb{R}^d$. Let $\boldsymbol{W} \in \mathbb{R}^{n \times d}$ be a random matrix such that each $W_{i,j}$ is an independent Gaussian random variable. Then, for every $\epsilon \in (0, 3)$ we have

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

Lemma: Concentration of χ^2 Variables

• Let $X_1, \cdots, X_k$ be k independent Gaussian random variables, i.e. $X_i \sim N(0,1)$. The distribution of the random variable X_i^2 is call χ^2 and the distribution of the random variable $Z = X_1^2 + \cdots + X_k^2$ is called χ_k^2. Clearly, $\mathbb{E}[X_i^2] = 1$ and $\mathbb{E}[Z] = k$. The following lemma states that X_k^2 is concentrated around its mean.

• **Lemma.** Let $Z \sim \chi_k^2$. Then, for all $\epsilon > 0$ we have

$$\mathbb{P}[Z \leq (1-\epsilon)k] \leq e^{-\epsilon^2 k/6},$$

and for all $\epsilon \in (0,3)$ we have

$$\mathbb{P}[Z \geq (1+\epsilon)k] \leq e^{-\epsilon^2 k/6}.$$

Finally, for all $\epsilon \in (0,3)$,

$$\mathbb{P}[(1-\epsilon)k \leq Z \leq (1+\epsilon)k] \geq 1 - 2e^{-\epsilon^2 k/6}.$$

Proof of Lemma 2.

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

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

• Let $\boldsymbol{w}_i$ be the i-the row of $\boldsymbol{W}$. The random variable $\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle$ ($\boldsymbol{x} \in \mathbb{R}^d$ is given) is a weighted sum of d independent normal random variables and therefore it is normally distributed with zero mean and variance $\sum_j x_j^2 = \|\boldsymbol{x}\|^2 = 1$ (the variance of $(\boldsymbol{w}_i)_i * x_i$ is x_i^2, also $var(r_1 + r_2) = var(r_1) + var(r_2)$, i.e., $\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle$ is the standard Gaussian).

• Therefore, the random variable $\|\boldsymbol{W}\boldsymbol{x}\|^2 = \sum_{i=1}^n (\langle \boldsymbol{w}_i, \boldsymbol{x}\rangle)^2$ has a χ_n^2 distribution. The claim now follows directly from a measure concentration property of χ^2 random variables.

Lemma 3. [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 at least $1 - \delta$ over a choice of a random matrix $\boldsymbol{W} \in \mathbb{R}^{n \times d}$ such that each element of $\boldsymbol{W}$ is distributed normally with zero mean and variance of $1/n$ we have

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

Proof. Combining Lemma 2 and the union bound ($P(A \bigcup B) \leq P(A) + P(B)$) we have that for every $\epsilon \in (0, 3)$:

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

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

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

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

Principal Component Analysis (PCA)

• Another natural criterion for choosing $\boldsymbol{W}$ is in a way that will enable a reasonable recovery of $\boldsymbol{x}$ from $\boldsymbol{Wx}$ is possible.

Principal component analysis (PCA)

• Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m$ be m vectors in $\mathbb{R}^d$. We would like to reduce the dimensionality of these vectors using a linear transformation.

• A matrix $\boldsymbol{W} \in \mathbb{R}^{n \times d}$, where $n < d$, induces a mapping $\boldsymbol{x} \rightarrow \boldsymbol{W}\boldsymbol{x}$, where $\boldsymbol{W}\boldsymbol{x} \in \mathbb{R}^n$ is the lower dimensionality representation of $\boldsymbol{x}$. Then, a second matrix $\boldsymbol{U} \in \mathbb{R}^{d \times n}$ can be used to (approximately) recover each original vector $\boldsymbol{x}$ from its compressed version.

• That is, for a compressed vector $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$, where $\boldsymbol{y}$ is in the low dimensional space $\mathbb{R}^n$, we can construct $\tilde{\boldsymbol{x}} = \boldsymbol{U}\boldsymbol{y}$, so that $\tilde{\boldsymbol{x}}$ is the recovered version of $\boldsymbol{x}$ and resides in the original high dimensional space $\mathbb{R}^d$. We want $\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|$ to be small.

• **PCA:** find the compression matrix $\boldsymbol{W}$ and the recovering matrix $\boldsymbol{U}$ so that the total squared distance between the original and recovered vectors is minimal:

$$\arg \min_{\boldsymbol{W} \in \mathbb{R}^{n \times d}, \boldsymbol{U} \in \mathbb{R}^{d \times n}} \sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{U}\boldsymbol{W}\boldsymbol{x}_i\|^2. \tag{1}$$

- To solve problem (1), we first show that the optimal solution takes a specific form.

- **Lemma 1.** Let $(\boldsymbol{U}, \boldsymbol{W})$ be a solution to (1). Then the columns of $\boldsymbol{U}$ are orthonormal (namely, $\boldsymbol{U}^\top \boldsymbol{U}$ is the identity matrix of $\mathbb{R}^n$) and $\boldsymbol{W} = \boldsymbol{U}^\top$.

Proof.

• Fix any $\boldsymbol{U}, \boldsymbol{W}$ and consider the mapping $\boldsymbol{x} \to \boldsymbol{UWx}$. The range of this mapping, $R = \{\boldsymbol{UWx} : \boldsymbol{x} \in \mathbb{R}^d\}$, is an n dimensional linear subspace of $\mathbb{R}^d$.

• Let $\boldsymbol{V} \in \mathbb{R}^{d \times n}$ be a matrix whose columns form an orthonormal basis of this subspace, namely, the range of $\boldsymbol{V}$ is R and $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}$. Therefore, each vector in R can be written as $\boldsymbol{Vy}$ where $\boldsymbol{y} \in \mathbb{R}^n$.

• Also, for every $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{y} \in \mathbb{R}^n$, we have

$$\|\boldsymbol{x} - \boldsymbol{Vy}\|^2 = \|\boldsymbol{x}\|^2 + \boldsymbol{y}^\top \boldsymbol{V}^\top \boldsymbol{Vy} - 2\boldsymbol{y}^\top \boldsymbol{V}^\top \boldsymbol{x} = \|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - 2\boldsymbol{y}^\top (\boldsymbol{V}^\top \boldsymbol{x}),$$

where we used the fact that $\boldsymbol{V}^\top \boldsymbol{V}$ is the identity matrix of $\mathbb{R}^n$.

• Minimizing the preceding expression w.r.t. $\boldsymbol{y}$ by comparing the gradient w.r.t. $\boldsymbol{y}$ to zero gives that $\boldsymbol{y} = \boldsymbol{V}^\top \boldsymbol{x}$. Therefore, for each $\boldsymbol{x}$ we have that

$$\boldsymbol{V}\boldsymbol{V}^\top \boldsymbol{x} = \arg\min_{\tilde{\boldsymbol{x}} \in R} \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|^2.$$

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

$$\sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{U}\boldsymbol{W}\boldsymbol{x}_i\|_2^2 \geq \sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{V}\boldsymbol{V}^\top \boldsymbol{x}_i\|_2^2.$$

Since this holds for every $\boldsymbol{U}, \boldsymbol{W}$ the proof of the lemma follows.

- Lemma 1 indicates that we can rewrite the optimization problem (1) as follows:

$$\arg \min_{\boldsymbol{U} \in \mathbb{R}^{d \times n}, \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}} \sum_{i=1}^{m} \|\boldsymbol{x}_i - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}_i\|_2^2. \tag{2}$$

- We further simplify the optimization problem: For every $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{U} \in \mathbb{R}^{d \times n}$ such that $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$ we have

$$\begin{aligned} \|\boldsymbol{x} - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x}\|^2 &= \|\boldsymbol{x}\|^2 - 2\boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x} \\ &= \|\boldsymbol{x}\|^2 - \boldsymbol{x}^\top \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{x} = \|\boldsymbol{x}\|^2 - \operatorname{trace}(\boldsymbol{U}^\top \boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{U}), \end{aligned} \tag{3}$$

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 (2) as follows:

$$\arg \max_{\boldsymbol{U} \in \mathbb{R}^{d \times n} : \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}} \operatorname{trace}\Big(\boldsymbol{U}^\top \sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{U}\Big). \tag{4}$$

- $\operatorname{trace}(\boldsymbol{A} + \boldsymbol{B}) = \operatorname{trace}(\boldsymbol{A}) + \operatorname{trace}(\boldsymbol{B})$

• Let $\boldsymbol{A} = \sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^\top$.

• The matrix $\boldsymbol{A}$ is symmetric (all eigenvalues are real and $\boldsymbol{A}$ is diagonalizable) and therefore it can be written using its spectral decomposition as $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top$, where $\boldsymbol{D}$ is diagonal and $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{I}$. Here, the elements on the diagonal of $\boldsymbol{D}$ are the eigenvalues of $\boldsymbol{A}$ and the columns of $\boldsymbol{V}$ are the corresponding eigenvectors.

• We assume without of generality that $D_{1,1} \geq D_{2,2} \geq \cdots \geq D_{d,d}$. Since $\boldsymbol{A}$ is positive semidefinite it holds that $D_{d,d} \geq 0$. We claim that the solution to (4) is the matrix $\boldsymbol{U}$ whose columns are the n eigenvectors of $\boldsymbol{A}$ corresponding to the largest n eigenvalues.

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

Proof.

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

$$\boldsymbol{U}^\top \boldsymbol{A}\boldsymbol{U} = \boldsymbol{B}^\top \boldsymbol{V}^\top \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{V}\boldsymbol{B} = \boldsymbol{B}^\top \boldsymbol{D}\boldsymbol{B},$$

and therefore

$$\mathrm{trace}(\boldsymbol{U}^\top \boldsymbol{A}\boldsymbol{U}) = \mathrm{trace}(\boldsymbol{B}^\top \boldsymbol{D}\boldsymbol{B}) = \sum_{j=1}^{d} D_{j,j} \sum_{i=1}^{n} B_{j,i}^2.$$

• Note that $\boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{U}^\top \boldsymbol{V}\boldsymbol{V}^\top \boldsymbol{U} = \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$. Therefore, the columns of $\boldsymbol{B}$ are also orthonormal, implying that $\sum_{i=1}^{n} \sum_{j=1}^{d} B_{j,i}^2 = \sum_{j=1}^{d} \sum_{i=1}^{n} B_{j,i}^2 = n$. In addition, let $\tilde{\boldsymbol{B}} \in \mathbb{R}^{d \times d}$ be a matrix such that its first n columns are the columns of $\boldsymbol{B}$ and in addition $\tilde{\boldsymbol{B}}^\top \tilde{\boldsymbol{B}} = \boldsymbol{I}$. Then, we have $\sum_{i=1}^{d} \tilde{B}_{j,i}^2 = 1, \forall j$, implying that $\sum_{i=1}^{n} B_{j,i}^2 \leq 1$.

• It follows that

$$\mathrm{trace}(\boldsymbol{U}^\top \boldsymbol{A}\boldsymbol{U}) = \sum_{j=1}^{d} D_{jj}\beta_j \le \max_{\beta\in[0,1]^d:\|\beta\|_1=n} \sum_{j=1}^{d} D_{j,j}\beta_j. \ \left(\beta_j = \sum_{i=1}^{n} B_{j,i}^2 \in [0,1],\ \sum_{j=1}^{d}\beta_j = n\right)$$

It is not hard to verify that the right-hand side equals to $\sum_{j=1}^{n} D_{j,j}$ (note that $D_{j,j}$ has been sorted and by setting $\beta_1 = \cdots \beta_n = 1$ and $\beta_{n+1} = \cdots = \beta_d = 0$).

• We have therefore shown that for every matrix $\boldsymbol{U} \in \mathbb{R}^{d\times n}$ with orthonormal columns it hold that $\mathrm{trace}(\boldsymbol{U}^\top \boldsymbol{A}\boldsymbol{U}) \le \sum_{j=1}^{n} D_{j,j}$.

• On the other hand, if we set $\boldsymbol{U}$ to be the matrix whose columns are the n leading eigenvectors of $\boldsymbol{A}$ we obtain that

$$\mathrm{trace}(\boldsymbol{U}^\top \boldsymbol{A}\boldsymbol{U}) = \sum_{j=1}^{n} D_{j,j}, \text{ using spectral decomposition of } \boldsymbol{A}.$$

Remark. The proof of the above theorem tells us that:

> the value of the objective of (4) is $\sum_{i=1}^{n} D_{i,i}$;

> combine this with (3) and note that $\sum_{i=1}^{m} \|\boldsymbol{x}_i\|^2 = \mathrm{trace}(\boldsymbol{A}) = \sum_{i=1}^{d} D_{i,i}$ we obtain that the optimal objective value of (1) is $\sum_{i=n+1}^{d} D_{i,i}$.

Remark. It is a common practice to "center" the examples before applying PCA. That is, we first calculate $\mu = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{x}_i$ and then apply PCA on the vectors $(\boldsymbol{x}_1 - \mu), \cdots, (\boldsymbol{x}_m - \mu)$.

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 $\boldsymbol{A}$) plus $O(md^2)$ (for constructing the matrix $\boldsymbol{A}$).

- How to calculate PCA solution efficiently when $d \gg m$?

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

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

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

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

PCA

Input A matrix of m examples $\boldsymbol{X} \in \mathbb{R}^{m \times d}$, number of components n.

if $(m > d)$

$\boldsymbol{A} = \boldsymbol{X}^\top \boldsymbol{X}$, let $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n$ be the eigenvectors of $\boldsymbol{A}$ with largest eigenvalues

else

$\boldsymbol{B} = \boldsymbol{X}\boldsymbol{X}^\top$

Let $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_n$ be the eigenvectors of $\boldsymbol{B}$ with largest eigenvalues for $i = 1, \cdots, n$ set $\boldsymbol{u}_i = \frac{1}{\|\boldsymbol{X}^\top \boldsymbol{v}_i\|} \boldsymbol{X}^\top \boldsymbol{v}_i$

output $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n$

• 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. Note that $\boldsymbol{A} = \sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i^\top \sim \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ with $\lambda_1 = 0$ ($\boldsymbol{v}_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$) and $\lambda_2 = 2$ ($\boldsymbol{v}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$).

• Suppose we apply PCA to this data. Then, the eigenvector corresponding to the largest eigenvalue is $(1/\sqrt{2}, 1/\sqrt{2})$. When projecting a point $(x, x+y)$ on this principal component we will obtain the scalar $(2x+y)/\sqrt{2}$. The reconstruction of the original vector will be $(1/\sqrt{2}, 1/\sqrt{2}) * (2x+y)/\sqrt{2} = ((x+y/2), (x+y/2))$. In Figure 1, we depict the original versus reconstructed data.

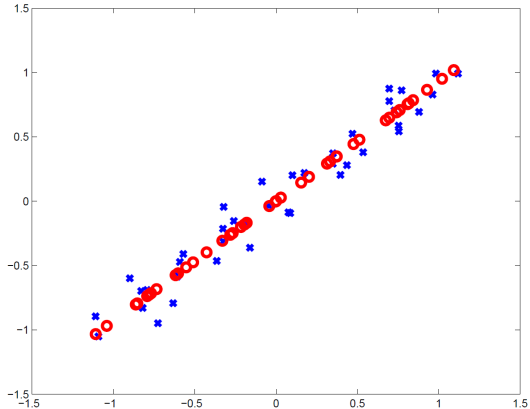
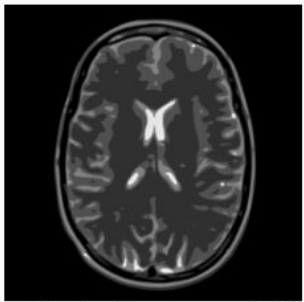


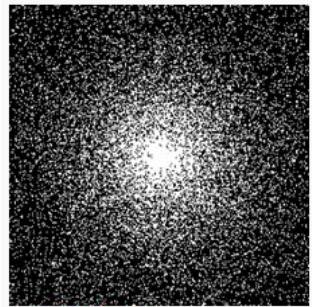
Figure: A set of vectors in $\mathbb{R}^2$ (blue) and their reconstruction after dimensionality reduction to $\mathbb{R}^1$ using PCA (red).

Compressed Sensing

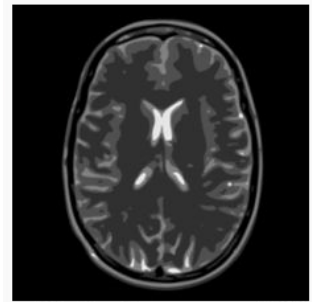
• Compressed sensing is a dimensionality reduction technique which utilizes a prior assumption that the original vector is sparse in some basis.



Ground truth



25% subsampling in k-space (MRI)



Reconstruction

• To motivate compressed sensing, consider a vector $\boldsymbol{x} \in \mathbb{R}^d$ that has at most $s \ll d$ nonzero elements. That is,

$$\|\boldsymbol{x}\|_0 := |\{i : x_i \neq 0\}| \leq s.$$

Clearly, we can compress $\boldsymbol{x}$ by representing it using s (index, value) pairs. This compression is lossless – we can reconstruct $\boldsymbol{x}$ exactly from the s (index, value) pairs.

• Now, let us take one step forward and assume that $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{\alpha}$ where $\boldsymbol{\alpha}$ is a sparse vector, $\|\boldsymbol{\alpha}\|_0 \leq s$, and $\boldsymbol{U}$ is a fixed orthonormal matrix. That is, $\boldsymbol{x}$ has a sparse representation in another basis. It turns out that many natural vectors are (at least approximately) sparse in some representations. – For instance, the natural image in wavelet representation.

• Can we still compress $\boldsymbol{x}$ into roughly s numbers?

• One simple way to do this is to multiply $\boldsymbol{x}$ by $\boldsymbol{U}^\top$, which yields the sparse vector α, and then represent α by its s (index,value) pairs. However, this requires us to first to "sense" $\boldsymbol{x}$, to store it, and then to multiply it by $\boldsymbol{U}^\top$.

• 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 $\boldsymbol{x}$ without losing information.

• The number of measurements needed is order of $s \log(d)$ rather than order d. That is, we roughly acquire only the important information about the signal.

• It is possible to reconstruct any sparse signal fully if it was compressed by $\boldsymbol{x} \rightarrow \boldsymbol{W}\boldsymbol{x}$, where $\boldsymbol{W}$ 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.

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

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

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

$$\tilde{\boldsymbol{x}} \in \arg \min_{\boldsymbol{v}: \boldsymbol{W}\boldsymbol{v} = \boldsymbol{y}} \|\boldsymbol{v}\|_0,$$

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

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

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

The reconstruction scheme given in Theorem 1 seems to be nonefficient because we need to minimize a combinatorial objective (the sparsity of $\boldsymbol{v}$). Quite surprisingly, it turns out that we can replace the combinatorial objective, $\|\boldsymbol{v}\|_0$, with a convex objective, $\|\boldsymbol{v}\|_1$, which leads to a linear programming problem that can be solved efficiently.

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

$$\boldsymbol{x} = \arg \min_{\boldsymbol{v}:\boldsymbol{W}\boldsymbol{v}=\boldsymbol{y}} \|\boldsymbol{v}\|_0 = \arg \min_{\boldsymbol{v}:\boldsymbol{W}\boldsymbol{v}=\boldsymbol{y}} \|\boldsymbol{v}\|_1.$$

When a matrix is RIP

- A random matrix with $n \geq \Omega(s \log(d))$ are likely to be RIP.

- In fact, multiplying a random matrix by an orthonormal matrix also provides an RIP matrix.

- This is important for compressing signals of the form $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{\alpha}$ where $\boldsymbol{x}$ is not sparse but $\boldsymbol{\alpha}$ is sparse. In that case, if $\boldsymbol{W}$ is a random matrix and we compress using $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ then this is the same as compressing $\boldsymbol{\alpha}$ by $\boldsymbol{y} = (\boldsymbol{W}\boldsymbol{U})\boldsymbol{\alpha}$ and since $\boldsymbol{W}\boldsymbol{U}$ is also RIP we can reconstruct $\boldsymbol{\alpha}$ (and thus also $\boldsymbol{x}$) from $\boldsymbol{y}$.

More precisely, we have:

Theorem 3. Let $\boldsymbol{U}$ be an arbitrary fixed $d \times d$ orthonormal matrix, let ϵ, δ be scalars on $(0, 1)$, 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 $\boldsymbol{W} \in \mathbb{R}^{n \times d}$ be a matrix s.t. each element of $\boldsymbol{W}$ is distributed normally with zero mean and variance of $1/n$. Then, with probability at least $1 - \delta$ over the choice of $\boldsymbol{W}$, the matrix $\boldsymbol{WU}$ is (ϵ, s)-RIP.