1. Review of matrix eigendecomposition

1.1. Eigenvalues and eigenvectors. Let $\mathbf{A}$ be an $n \times n$ real matrix (this is often denoted as $\mathbf{A} \in \mathbb{R}^{n\times n}$). The characteristic polynomial of $\mathbf{A}$ is

$$p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = \prod(\lambda - \lambda_i)^{a_i}.$$

The (complex) roots λ_i of the characteristic equation $p(\lambda) = 0$ are called the eigenvalues of $\mathbf{A}$. For a specific eigenvalue λ_i of $\mathbf{A}$, any nonzero vector $\mathbf{v}_i$ satisfying

$$(\mathbf{A} - \lambda_i\mathbf{I})\mathbf{v}_i = \mathbf{0}$$

or equivalently,

$$\mathbf{A}\mathbf{v}_i = \lambda_i\mathbf{v}_i$$

is called an eigenvector of $\mathbf{A}$ associated to λ_i. All eigenvectors associated to an eigenvalue λ_i together with the zero vector $\mathbf{0}$ form a subspace, called the eigenspace; it is denoted as $\mathrm{E}(\lambda_i) = \mathrm{N}(\mathbf{A} - \lambda_i\mathbf{I})$. The dimension g_i of $\mathrm{E}(\lambda_i)$ is called the geometric multiplicity of λ_i, while the degree a_i of the factor $(\lambda - \lambda_i)^{a_i}$ in $p(\lambda)$ is called the algebraic multiplicity of λ_i. Note that we must have $\sum a_i = n$ and for all i, $1 \le g_i \le a_i$.

Example 1.1. Let

$$\mathbf{A} = \begin{pmatrix} 3 & 0 & 0 \\ 5 & 1 & -1 \\ -2 & 2 & 4 \end{pmatrix}.$$

Find all the above quantities.
Answer. The eigenvalues are $\lambda_1 = 3, \lambda_2 = 2$ with $a_1 = 2, a_2 = 1$ and $g_1 = g_2 = 1$. The corresponding eigenvectors are $\mathbf{v}_1 = (0, 1, -2)^T, \mathbf{v}_2 = (0, 1, -1)^T$.

Theorem 1.1. *Let $\mathbf{A}$ be a real square matrix whose eigenvalues are $\lambda_1, \ldots, \lambda_n$ (counting multiplicities). Then*

$$\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i \qquad \textit{and} \quad \operatorname{trace}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i.$$

Definition 1.1. A square matrix $\mathbf{A}$ is diagonalizable if it is similar to a diagonal matrix, i.e., there exist an invertible matrix $\mathbf{P}$ and a diagonal matrix Λ such that

$$\mathbf{A} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{-1}.$$

Remark. The above equation implies that $\mathbf{A}\mathbf{p}_i = \lambda_i\mathbf{p}_i$ for $1 \le i \le n$, where $\mathbf{p}_i$ are the columns of $\mathbf{P}$. This shows that the λ_i are the eigenvalues of $\mathbf{A}$ and $\mathbf{p}_i$ the associated eigenvectors. Thus, the above factorization is called the eigenvalue decomposition of $\mathbf{A}$.

Example 1.2. The matrix

$$\mathbf{A} = \begin{pmatrix} 0 & 1 \\ 3 & 2 \end{pmatrix}$$

is diagonalizable because

$$\begin{pmatrix} 0 & 1 \\ 3 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 3 & -1 \end{pmatrix} \begin{pmatrix} 3 & \\ & -1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 3 & -1 \end{pmatrix}^{-1}$$

but $\mathbf{B} = \begin{pmatrix} 0 & 1 \\ -1 & 2 \end{pmatrix}$ is not.

Theorem 1.2. *A matrix* $\mathbf{A}$ *is diagonalizable if and only if it has* n *linearly independent eigenvectors.*

Corollary 1.3. *The following matrices are diagonalizable:*

- *Any matrix whose eigenvalues all have identical geometric and algebraic multiplicities, i.e.,* $g_i = a_i$ *for all* i*;*
- *Any matrix with* n *distinct eigenvalues;*

In next section, we show that symmetric matrices are always diagonalizable.

1.2. Symmetric matrices. Recall that an orthogonal matrix is a square matrix whose columns and rows are both orthogonal unit vectors (i.e., orthonormal vectors):

$$\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbf{I},$$

or equivalently,

$$\mathbf{Q}^{-1} = \mathbf{Q}^T.$$

Theorem 1.4. *Let* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *be a symmetric matrix. Then*

- *All the eigenvalues of* $\mathbf{A}$ *are real;*
- $\mathbf{A}$ *is orthogonally diagonalizable, i.e., there exists an orthogonal matrix* $\mathbf{Q}$ *and a diagonal matrix* Λ *such that*

$$\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^T.$$

Proof. This theorem can be proved by induction. □

Remark.

- For symmetric matrices, the eigenvalue decomposition is also called the spectral decomposition.
- The converse is also true. Therefore, a matrix is symmetric if and only if it is orthogonally diagonalizable.
- Write $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ and $\mathbf{Q} = [\mathbf{q}_1, \dots, \mathbf{q}_n]$. Then the product can be expanded as

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{q}_i \mathbf{q}_i^T.$$

- We often sort the diagonals of Λ in decreasing order such that

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n.$$

Example 1.3. Find the spectral decomposition of the following matrix

$$\mathbf{A} = \begin{pmatrix} 0 & 2 \\ 2 & 3 \end{pmatrix}$$

Answer.

$$\mathbf{A} = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 4 & \\ & -1 \end{pmatrix} \cdot \frac{1}{\sqrt{5}} \begin{pmatrix} 1 & -2 \\ 2 & 1 \end{pmatrix}^T$$

Definition 1.2. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is positive semidefinite if $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$. It is positive definite if $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ whenever $\mathbf{x} \neq \mathbf{0}$.

Theorem 1.5. *Let $\mathbf{A}$ be a symmetric matrix. It is positive definite (semidefinite) if and only if all the eigenvalues are positive (nonnegative).*

2. Singular Value Decomposition

2.1. Matrix representation of high dimensional data. High dimensional data exists in various forms, such as images, videos, hyperspectral images, audio signals, and text documents. All of them are commonly represented as points in high dimensional Euclidean spaces $\mathbb{R}^d$. We often store them as $n \times d$ matrices $\mathbf{X} \in \mathbb{R}^{n\times d}$ or $d \times n$ matrices $\mathbf{Y} \in \mathbb{R}^{d\times n}$, whichever is more convenient for the specific task.

2.2. Singular value decomposition (SVD). Suppose $\mathbf{X} \in \mathbb{R}^{n\times d}$ (with rows representing points). Let $\mathbf{C} = \mathbf{X}^T\mathbf{X} \in \mathbb{R}^{d\times d}$.

Proposition 2.1. *The matrix $\mathbf{C}$ is positive semidefinite.*

We apply the spectral decomposition theorem to $\mathbf{C}$ to derive the singular value decomposition of $\mathbf{X}$. First, there exists an orthogonal matrix $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_d] \in \mathbb{R}^{d\times d}$ and a diagonal matrix $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d) \in \mathbb{R}^{d\times d}$ with $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$ such that

$$\mathbf{C} = \mathbf{X}^T\mathbf{X} = \mathbf{V}\Lambda\mathbf{V}^T.$$

Rewrite the above equation as

$$\mathbf{X}^T\mathbf{X}\mathbf{V} = \mathbf{V}\Lambda.$$

Consider, for each $1 \leq i \leq d$, the ith column

$$\mathbf{X}^T\mathbf{X}\mathbf{v}_i = \lambda_i\mathbf{v}_i = \sigma_i^2\mathbf{v}_i, \tag{1}$$

where $\sigma_i = \sqrt{\lambda_i}$. For all $\sigma_1 \geq \cdots \geq \sigma_r > 0$, where $r = \text{rank}(\mathbf{C}) = \text{rank}(\mathbf{X})$, define

$$\mathbf{u}_i = \frac{1}{\sigma_i}\mathbf{X}\mathbf{v}_i \in \mathbb{R}^n.$$

Claim: $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are orthonormal vectors. The above is equivalent to

$$\mathbf{X}\mathbf{v}_i = \sigma_i\mathbf{u}_i, \quad i = 1, \ldots, r.$$

For all $r < i \leq n$ select unit vectors $\mathbf{u}_i \in \mathbb{R}^n$ such that

$$\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_r, \mathbf{u}_{r+1}, \ldots, \mathbf{u}_n] \in \mathbb{R}^{n\times n}$$

is an orthogonal matrix. Let Σ be an $n \times d$ matrix whose entries are all zero except the top $r \times r$ block

$$\Sigma(1:r, 1:r) = \text{diag}(\sigma_1, \ldots, \sigma_r).$$

It is easy to verify that with the above choices of $\mathbf{U}$ and Σ, we must have

$$\mathbf{X}\mathbf{V} = \mathbf{U}\Sigma$$

Therefore, we have proved the following result.

Theorem 2.2. *For any matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, there exist orthogonal matrices $\mathbf{U} \in \mathbb{R}^{n\times n}, \mathbf{V} \in \mathbb{R}^{d\times d}$ and a diagonal matrix $\Sigma \in \mathbb{R}^{n\times d}$ (with nonnegative entries) such that*

$$\mathbf{X}_{n\times d} = \mathbf{U}_{n\times n}\Sigma_{n\times d}\mathbf{V}_{d\times d}^T.$$

Definition 2.1. The above decomposition of a given matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ is called the Singular Value Decomposition (SVD) of $\mathbf{X}$. The diagonals of Σ (including zero) are called the singular values of $\mathbf{X}$; the columns of $\mathbf{U}, \mathbf{V}$ are called the left and right singular vectors, respectively.

Remark. The above decomposition is often called the full SVD of $\mathbf{X}$, to distinguish from other versions:

- Economic/compact SVD: Let $r = \operatorname{rank}(\mathbf{X})$. Define

$$\begin{aligned}\mathbf{U}_{n\times r} &= [\mathbf{u}_1, \dots, \mathbf{u}_r] \in \mathbb{R}^{n\times r}\\ \mathbf{V}_{d\times r} &= [\mathbf{v}_1, \dots, \mathbf{v}_r] \in \mathbb{R}^{d\times r}\\ \Sigma_{r\times r} &= \operatorname{diag}(\sigma_1, \dots, \sigma_r) \in \mathbb{R}^{r\times r}\end{aligned}$$

  We then have

$$\mathbf{X} = \mathbf{U}_{n\times r}\Sigma_{r\times r}\mathbf{V}_{d\times r}^T.$$

- Rank-1 decomposition:

$$\mathbf{X} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T.$$

  This has the interpretation that $\mathbf{X}$ is a weighted sum of rank-one matrices.

In sum, $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$ where both $\mathbf{U}, \mathbf{V}$ have orthonormal columns and Σ is diagonal. Furthermore, $\mathbf{X}^T = \mathbf{V}\Sigma^T\mathbf{U}^T$ is the SVD of $\mathbf{X}^T$.
Remark. For any version, the SVD of a matrix is not unique.

Example 2.1. Compute the SVD of

$$\mathbf{X} = \begin{pmatrix} 1 & -1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Answer.

$$\mathbf{X} = \begin{pmatrix} \frac{2}{\sqrt{6}} & 0 \\ -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \end{pmatrix} \cdot \begin{pmatrix} \sqrt{3} & \\ & 1 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}^T$$

2.3. Low-rank approximation of matrices. Recall that a norm associated with a vector space $\mathcal{V}$ is a function $\|\cdot\| : \mathcal{V} \to \mathbb{R}$ that satisfies three conditions:

- $\|\mathbf{v}\| \geq 0$ for all $\mathbf{v} \in \mathcal{V}$ and $\|\mathbf{v}\| = 0$ iff $\mathbf{v} = \mathbf{0}$
- $\|k\mathbf{v}\| = |k|\|\mathbf{v}\|$
- $\|\mathbf{v}_1 + \mathbf{v}_2\| \leq \|\mathbf{v}_1\| + \|\mathbf{v}_2\|$ for all $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{V}$

Example 2.2. In $\mathbb{R}^d$, there are at least three different norms:

- 2-norm (or Euclidean norm): $\|\mathbf{x}\|_2 = \sqrt{\sum x_i^2} = \sqrt{\mathbf{x}^T\mathbf{x}}$

- 1-norm (Taxicab norm or Manhattan norm): $\|\mathbf{x}\|_1 = \sum |x_i|$
- ∞-norm (maximum norm): $\|\mathbf{x}\|_\infty = \max |x_i|$

When unspecified, it is defaulted the Euclidean norm.

We next define matrix norms. Just like vector norm is used to measure the magnitude of vectors ($\|\mathbf{v}\|$) and quantify the distance between vectors ($\|\mathbf{u} - \mathbf{v}\|$), matrix norm is used similarly.

Definition 2.2. The Frobenius norm of a matrix is defined as

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$$

Example 2.3. In the last example, $\|\mathbf{X}\|_F = 2$.

Proposition 2.3.

$$\|\mathbf{A}\|_F^2 = \text{trace}(\mathbf{A}^T\mathbf{A}) = \text{trace}(\mathbf{A}\mathbf{A}^T)$$

Theorem 2.4. *For any matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$,*

$$\|\mathbf{A}\|_F^2 = \sum \sigma_i^2$$

A second matrix norm is the 2-norm, or the spectral norm.

Definition 2.3. The spectral norm of a matrix is defined as

$$\|\mathbf{A}\|_2 = \max_{\mathbf{q}\in\mathbb{R}^d : \|\mathbf{q}\|_2 = 1} \|\mathbf{A}\mathbf{q}\|_2$$

Theorem 2.5. *For any matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, a maximizer of the above problem is the first right singular vector $\mathbf{v}_1$ of $\mathbf{A}$ and the maximum value is*

$$\|\mathbf{A}\|_2 = \sigma_1.$$

Proof. Consider the full SVD of $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$. For any unit vector $\mathbf{q} \in \mathbb{R}^d$, write $\mathbf{q} = \mathbf{V}\alpha$ for some unit vector $\alpha \in \mathbb{R}^d$. Then $\mathbf{A}\mathbf{q} = \mathbf{A}(\mathbf{V}\alpha) = \mathbf{U}\Sigma\alpha$. Accordingly, $\|\mathbf{A}\mathbf{q}\|_2 = \|\mathbf{U}\Sigma\alpha\|_2 = \|\Sigma\alpha\|_2 = \sqrt{\sum \sigma_i^2\alpha_i^2} \leq \sigma_1$, where the equality holds when $\alpha = \pm\mathbf{e}_1$ and correspondingly, $\mathbf{y} = \pm\mathbf{V}\mathbf{e}_1 = \pm\mathbf{v}_1$. □

Example 2.4. In the last example, $\|\mathbf{X}\|_2 = \sqrt{3}$.

Corollary 2.6. *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$. Then for all $\mathbf{x} \in \mathbb{R}^d$,*

$$\|\mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{A}\|_2\|\mathbf{x}\|_2 = \sigma_1\|\mathbf{x}\|_2.$$

We note that the Frobenius and spectral norms of a matrix correspond to the 2- and ∞-norms of the vector of singular values. The 1-norm of singular values is called the nuclear norm of $\mathbf{A}$.

Definition 2.4. The nuclear norm of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is

$$\|\mathbf{A}\|_* = \sum \sigma_i.$$

Example 2.5. In the last example, $\|\mathbf{X}\|_* = \sqrt{3} + 1$.

We now consider the low rank matrix approximation problem.

Definition 2.5. For any $1 \leq k \leq r$, define $\mathbf{A}_k = \sum_{i=1}^k \sigma_i\mathbf{u}_i\mathbf{v}_i^T \in \mathbb{R}^{n\times d}$ as the truncated svd.

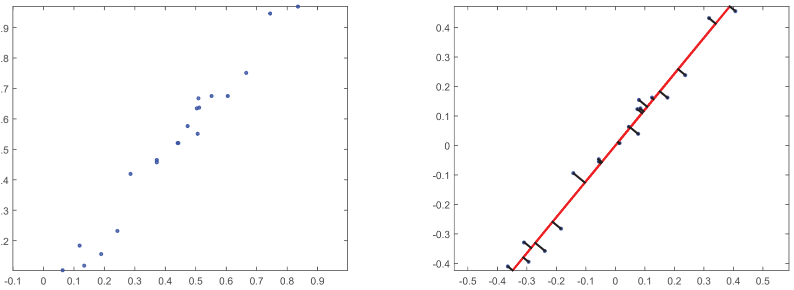

Figure 2. Illustration of the PCA fitting problem

Remark. Clearly, $\operatorname{rank}(\mathbf{A}_k) = k$.

Theorem 2.7. *For each $1 \le k \le r$, $\mathbf{A}_k$ is the best rank-k approximation to $\mathbf{A}$ under the Frobenius norm:*

$$\min_{\mathbf{B}\,:\,\operatorname{rank}\mathbf{B}=k} \|\mathbf{A}-\mathbf{B}\|_F = \|\mathbf{A}-\mathbf{A}_k\|_F = \sqrt{\sum_{i>k} \sigma_i^2}.$$

Proof. □

Remark.

- The theorem still holds true if the spectral norm is used instead. In this case, the minimum value is σ_{k+1}.
- The constraint $\operatorname{rank}\mathbf{B} = k$ can be changed to $\operatorname{rank}\mathbf{B} \le k$ without changing the solution.

Example 2.6. In the last example, the best rank-1 approximation (under the Frobenius/spectral norm) is

$$\mathbf{X}_1 = \begin{pmatrix} 1 & -1 \\ -\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{pmatrix}.$$

Finally, we also mention the matrix 1- and ∞-norms:

$$\|\mathbf{A}\|_1 = \max_j \sum_i |a_{ij}| \quad \text{(maximum absolute column sum)}$$

$$\|\mathbf{A}\|_\infty = \max_i \sum_j |a_{ij}| \quad \text{(maximum absolute row sum)}$$

2.4. Orthogonal Best-Fit Subspace. Consider the following problem: given n points $\mathbf{x}_i \in \mathbb{R}^d$, find the "best-fit" k-dimensional subspace (see Fig. 2) which minimizes

$$\sum \|\mathbf{x}_i - \mathcal{P}_S(\mathbf{x}_i)\|_2^2$$

Remark. Compare with the least squares fitting problem.

Let $\mathbf{m} \in \mathbb{R}^d$ represent a fixed point and $\mathbf{B} \in \mathbb{R}^{d\times k}$ an orthonormal basis of $\mathcal{S}$ (i.e., $\mathbf{B}^T\mathbf{B} = \mathbf{I}_{k\times k}$, but not $\mathbf{B}\mathbf{B}^T = \mathbf{I}_{d\times d}$ for $k < d$) so that a parametric equation for the plane is

$$\mathbf{x} = \mathbf{m} + \mathbf{B}\alpha.$$

First,

$$\mathcal{P}_S(\mathbf{x}_i) = \mathbf{m} + \mathbf{B}\mathbf{B}^T(\mathbf{x}_i - \mathbf{m}).$$

We then rewrite the above problem as

$$\min_{\mathbf{m},\mathbf{B}} \sum \|\mathbf{x}_i - (\mathbf{m} + \mathbf{B}\mathbf{B}^T(\mathbf{x}_i - \mathbf{m}))\|^2$$

Using multivariable calculus, we can show that a best $\mathbf{m}$ is $\bar{\mathbf{x}} = \frac{1}{n}\sum \mathbf{x}_i$. Plugging in $\bar{\mathbf{x}}$ for $\mathbf{m}$ and letting $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}$ we rewrite the above equation as follows:

$$\min_{\mathbf{B}} \sum \|\tilde{\mathbf{x}}_i - \mathbf{B}\mathbf{B}^T\tilde{\mathbf{x}}_i\|^2.$$

In matrix notation, this is

$$\min_{\mathbf{B}} \|\widetilde{\mathbf{X}} - \widetilde{\mathbf{X}}\mathbf{B}\mathbf{B}^T\|_F^2$$

where $\widetilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n]^T \in \mathbb{R}^{n\times d}$. The minimum occurs when $\widetilde{\mathbf{X}}\mathbf{B}\mathbf{B}^T = \widetilde{\mathbf{X}}_k$, the best rank-$k$ approximation of $\widetilde{\mathbf{X}}$, and the corresponding minimizer $\mathbf{B}$ can be taken to be the matrix consisting of the top k right singular vectors of $\widetilde{\mathbf{X}}$:

$$\mathbf{B} = (\mathbf{v}_1, \ldots, \mathbf{v}_k) \qquad \text{where} \quad \widetilde{\mathbf{X}} = \mathbf{U}\Sigma\mathbf{V}^T.$$

We have thus proved the following result.

Theorem 2.8. *A best-fit subspace to the data is given by*

$$\mathbf{x} = \mathbf{m} + \mathbf{B}\alpha$$

where

$$\mathbf{m} = \textit{center}, \quad \mathbf{B} = \textit{top } k \textit{ right singular vectors of } \widetilde{\mathbf{X}},$$

and the projection of $\widetilde{\mathbf{X}}$ onto the best-fit k-plane is $\widetilde{\mathbf{X}}_k$.

Example 2.7. Computer demonstration.

2.5. Data analysis.

2.5.1. *Principal component analysis (PCA).*

Example 2.8. Consider a plane embedded in $\mathbb{R}^{10}$, from which we sample 100 points and add noise to the data. We can use PCA to reduce dimension and visualize the data. Additional effects include denoising.

Definition 2.6. The new coordinates of the centered data $\widetilde{\mathbf{X}}$ with respect to the basis $\mathbf{V}(:, 1:k)$, i.e., rows of

$$\widetilde{\mathbf{X}}\mathbf{V}(:, 1:k)^T = \mathbf{U}(:, 1:k)\Sigma(1:k, 1:k)$$

are called the principal components.

Theorem 2.9. *For each $1 \le j \le k$, the variance of the projection of $\tilde{\mathbf{X}}$ onto the $\mathbf{v}_j$ is σ_j^2. Moreover, these variances are the largest possible.*

Proof. We can show that the right singular vectors are solutions of the following problems:

$$\sigma_1 = \max_{\mathbf{v}:\, \|\mathbf{v}\|_2=1} \|\mathbf{A}\mathbf{v}\|_2$$
$$\sigma_2 = \max_{\mathbf{v}:\, \|\mathbf{v}\|_2=1, \mathbf{v}_1^T\mathbf{v}=0} \|\mathbf{A}\mathbf{v}\|_2$$
$$\sigma_3 = \max_{\mathbf{v}:\, \|\mathbf{v}\|_2=1, \mathbf{v}_i^T\mathbf{v}=0, i=1,2} \|\mathbf{A}\mathbf{v}\|_2$$
$$\vdots$$

□

Remark. PCA selects the orthogonal directions that maximize the variances of the projections onto each of those directions.

Algorithm 1 Principal component analysis (PCA)

```
Input: Data set $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$ and target dimension $k$
Return:
1: Center of the data set: $\bar{\mathbf{x}}$ (a point on the best-fit subspace)
2: Top $k$ right singular vectors of $\widetilde{\mathbf{X}}$ (an orthonormal basis for the best-fit subspace)
3: The singular values $\sigma_i$ (standard deviation of projection onto the $i$th principal direction)
4: Principal components $\widetilde{\mathbf{X}}\mathbf{V}(:, 1:k) = \mathbf{U}(:, 1:k)\Sigma(1:k, 1:k)$
```

Example 2.9 (Eigenfaces). Take many images of the facial images in frontal pose from the Exendend Yale B dataset and apply SVD to learn a basis.

Question 1: How do we select k?
There are several ways to determine k:

- Set $k=$ # dorminant singular values (effective rank)
- Choose k such that the top k principal directions explain a certain amount of variance in the data (e.g., 95%):

$$\frac{\sum_{i=1}^k \sigma_i^2}{\sum_i \sigma_i^2} > 95\%$$

Each criterion corresponds to a plot.

2.5.2. *Data compression.* Storage is reduced from nd to $k(n+d+1)+1$.

$$\mathbf{X} \approx \mathbf{U}\Sigma\mathbf{V}^T + \mathbf{m}$$

Example 2.10. Take a digital image (matrix) and apply SVD to obtain low-rank approximations. Show the corresponding images.

2.5.3. *Other practical issues.* **Question 2: Is SVD robust to outliers?**
We will do an experiment to find out.

Question 3: What if we have nonlinear data?

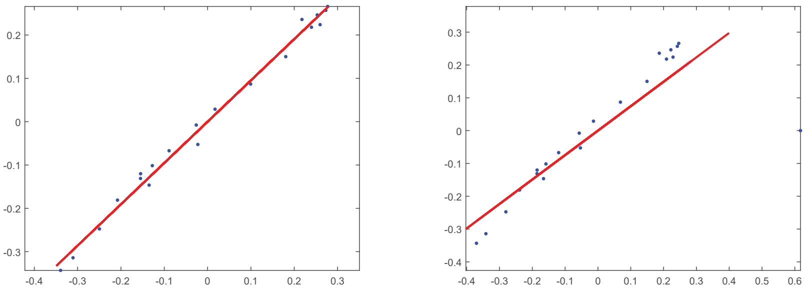

FIGURE 3. Sensitivity of PCA to outliers

SVD: A Summary

A. Singular Value Decomposition: $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T = \sum_i \sigma_i \mathbf{u}_i \mathbf{v}_i^T$. Here, $\mathbf{U}, \mathbf{V}$ have orthonormal columns and Σ is diagonal.

B. Matrix Norms.

- Frobenius: $\|\mathbf{A}\|_F = \sqrt{\sum a_{ij}^2} = \sqrt{\sum \sigma_i^2(\mathbf{A})}$
- Spectral: $\|\mathbf{A}\|_2 = \max_{\mathbf{q}:\|\mathbf{q}\|_2=1} \|\mathbf{A}\mathbf{q}\|_2 = \sigma_{\max}(\mathbf{A}) = \sigma_1(\mathbf{A})$
- Nuclear: $\|\mathbf{A}\|_* = \sum \sigma_i(\mathbf{A})$

C. Low Rank Matrix Approximation. The best rank-k approximation of a matrix $\mathbf{A}$ (under both the Frobenius norm and the spectral norm) is $\mathbf{A}_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^T$.

D. Principal Component Analysis (PCA).

$$\widetilde{\mathbf{X}} = \mathbf{X} - \bar{\mathbf{x}} = \mathbf{U}\Sigma\mathbf{V}^T$$

Things to keep in mind:

(1) The rows of $\mathbf{X}$ represent the given data points, while those of $\widetilde{\mathbf{X}}$ represent centered data.

(2) $\bar{\mathbf{x}}$ is the center of the data set $\mathbf{X}$, which always lies on the best-fit k-dimensional subspace (that minimizes the total squared orthogonal error).

(3) $\mathbf{V}(:, 1:k)$ is an orthonormal basis for the best-fit k-dimensional subspace.

(4) The rows of $\widetilde{\mathbf{X}}_k = \mathbf{U}(:, 1:k)\Sigma(1:k, 1:k)\mathbf{V}(:, 1:k)^T$ represent the coordinates of the projections of the centered data onto the best-fit subspace.

(5) The rows of $\mathbf{U}(:, 1:k)\Sigma(1:k, 1:k)$, which also equals $\widetilde{\mathbf{X}}\mathbf{V}(:, 1:k)$, are called the top k principal components of the data, being the coordinates of the projections on the best-fit subspace relative to the basis $\mathbf{V}(:, 1:k)$.

(6) The right singular vectors (i.e., columns of $\mathbf{V}$) are the principal directions in the data along which the variance of the projections onto each such direction is as large as possible (and equals the corresponding singular value squared, σ_j^2)

(7) The number of nonzero singular values is the matrix rank of $\widetilde{\mathbf{X}}$, while the number of "dominant" singular values is the "effective" rank of $\widetilde{\mathbf{X}}$.

In sum, PCA finds in a given data set low dimensional subspaces that

- *minimize* the total squared orthogonal error; and
- *maximize* the variances of the projections; and
- *preserve* the pairwise distances of the points in the data set as closely as possible.

Applications of SVD.

- Low-rank matrix approximation
- Subspace fitting
- Data compression (including denoising, dimensionality reduction, visualization)
- Much more: computing matrix pseudoinverse, solving redundant linear systems, etc.