Note 5: SVD and PCA

Math 198: Math for Machine Learning

1 Adjoints

How are a matrix $\mathbf{A}$ and its transpose $\mathbf{A}^\top$ related? Note that $\langle \mathbf{Ax}, \mathbf{y} \rangle = (\mathbf{Ax})^\top \mathbf{y} = \mathbf{x}^\top \mathbf{A}^\top \mathbf{y} = \langle \mathbf{x}, \mathbf{A}^\top \mathbf{y} \rangle$. So, when taking inner products, it seems that $\mathbf{A}^\top$ represents the action of $\mathbf{A}$ on the opposite argument. In fact, this relationship connects back to the underlying linear map T represented by $\mathbf{A}$. We define the *adjoint* of T, T^*, to be the linear map represented by $\mathbf{A}^\top$. Equivalently, $B(T(v), w) = C(v, T^*(w))$ for any appropriate *nondegenerate bilinear forms* B, C. (We will not define this term, as it is very far outside the scope of the class. Suffice to say that the inner products $\langle \cdot, \cdot \rangle_{\text{range}(T)}, \langle \cdot, \cdot \rangle_{\text{range}(T^*)}$ can be filled in for B and C, although inner products are not the only examples of nondegenerate bilinear forms.)[1]

We now explore how the adjoint T^* connects back to $\mathbf{A}$. Observe that any vector of the form $\mathbf{Av}$ is a linear combination of the columns of $\mathbf{A}$. Likewise, any vector of the form $\mathbf{A}^\top \mathbf{w}$ is a linear combination of the rows of $\mathbf{A}$. Therefore, $\text{Im}(T) = \text{range}(\mathbf{A}) = \text{col}(\mathbf{A})$, the *column space* of $\mathbf{A}$; $\text{Im}(T^*) = \text{range}(\mathbf{A}^\top) = \text{row}(\mathbf{A})$, the *row space* of $\mathbf{A}$.

For any linear map $T : \mathbb{R}^n \to \mathbb{R}^m$ (or the associated matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$), we therefore have the following four "fundamental subspaces" associated with T:

1. $\text{Im}(T)$, a.k.a. $\text{col}(\mathbf{A})$
2. $\ker(T)$, a.k.a. $\text{null}(\mathbf{A})$
3. $\text{Im}(T^*)$, a.k.a. $\text{row}(\mathbf{A})$
4. $\ker(T^*)$, a.k.a. $\text{null}(\mathbf{A}^\top)$.

2 Fundamental Theorem of Linear Algebra and SVD

It turns out these subspaces are related in a way captured by the first part of the Fundamental Theorem of Linear Algebra:

Theorem (FTLA, Part I)

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$. Then

(a) $\mathbb{R}^m = \text{range}(\mathbf{A}) \oplus \ker(\mathbf{A}^\top)$.

(b) $\text{rank}(\mathbf{A}) + \dim \ker(\mathbf{A}) = n$.

[1] If you want to learn more about adjoints, make sure you can solve all of Q4 on Homework 2, as understanding dual spaces is essential to understanding adjoints.

Proof

Proving (a) amounts to showing that $\ker(\mathbf{A}^\top) = \text{range}(\mathbf{A})^\perp$. We have

$$\begin{aligned}
\mathbf{x} \in \ker(\mathbf{A}^\top) &\iff \mathbf{A}^\top \mathbf{x} = \mathbf{0} \\
&\iff \mathbf{a}_i^\top \mathbf{x} = 0 \text{ for all } i = 1, \ldots, n \text{ (where } \mathbf{a}_i \text{ is the } i\text{'th column of } \mathbf{A}) \\
&\iff \mathbf{v}^\top \mathbf{x} = 0 \text{ for all } \mathbf{v} \in \text{range}(\mathbf{A}) \\
&\iff \mathbf{x} \in \text{Im}(\mathbf{A})^\perp.
\end{aligned}$$

Part (b) hinges on the fact that $\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{A}^\top)$. Assuming that fact, apply (a) combined with that $\dim(U \oplus V) = \dim U + \dim V$. □

Part (b) is usually known as the *Rank-Nullity Theorem*.

Singular Value Decomposition

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$, the FTLA gives us two natural-looking orthogonal decompositions involving the four fundamental subspaces of $\mathbf{A}$:

(i) $\mathbb{R}^n = \ker(\mathbf{A}) \oplus \text{range}(\mathbf{A}^\top)$

(ii) $\mathbb{R}^m = \ker(\mathbf{A}^\top) \oplus \text{range}(\mathbf{A})$

To dig deeper, we must examine the matrices $\mathbf{A}^\top\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{A}\mathbf{A}^\top \in \mathbb{R}^{m \times m}$. From an intuitive perspective, note that $(\mathbf{A}^\top\mathbf{A})_{i,j} = \mathbf{a}_i^\top \mathbf{a}_j$, where $\mathbf{a}_i$ is the i'th column of $\mathbf{A}$. Thus, $\mathbf{A}^\top\mathbf{A}$ gives some measure of the similarity between the columns of $\mathbf{A}$. Similarly, $\mathbf{A}\mathbf{A}^\top$ measures similarity between the rows of $\mathbf{A}$.

Lemma 1. If $\mathbf{A}$ has full column rank (i.e. $\mathbf{A}$ has n linearly independent columns), then $\mathbf{A}^\top\mathbf{A}$ is invertible.

Proof. Let $\mathbf{x} \in \ker(\mathbf{A}^\top\mathbf{A})$, i.e. $\mathbf{A}^\top\mathbf{A}\mathbf{x} = \mathbf{0}$. Then $\mathbf{A}\mathbf{x} \in \ker(\mathbf{A}^\top)$. By FTLA, $\mathbf{A}\mathbf{x} \in \text{range}(\mathbf{A})^\perp$. But clearly $\mathbf{A}\mathbf{x} \in \text{range}(\mathbf{A})$ as well, so $\mathbf{A}\mathbf{x} = \mathbf{0}$ by the fact that $\text{range}(\mathbf{A}) \perp \text{range}(\mathbf{A})^\perp$. Finally, since $\mathbf{A}$ is full-rank, $\mathbf{A}\mathbf{x} = \mathbf{0}$ implies that $\mathbf{x} = \mathbf{0}$. Thus, $\mathbf{A}^\top\mathbf{A}$ is a square matrix with trivial kernel, so it is invertible. □

Lemma 2. $\ker(\mathbf{A}^\top\mathbf{A}) = \ker(\mathbf{A})$, so $\text{rank}(\mathbf{A}^\top\mathbf{A}) = \text{rank}(\mathbf{A})$.

Proof. Exercise.

Lemma 3. $\mathbf{A}^\top\mathbf{A}$ is positive semi-definite (PSD).

Proof. Clearly $(\mathbf{A}^\top\mathbf{A})^\top = \mathbf{A}^\top\mathbf{A}$, so it is symmetric. To show that it is PSD, see that $\mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x} = \langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{x} \rangle \geq 0$ with equality iff $\mathbf{A}\mathbf{x} = 0$. □

To sum up what we know about $\mathbf{A}^\top\mathbf{A}$:

(i) It preserves the kernel and rank of $\mathbf{A}$

(ii) It is PSD.

If we take $\mathbf{A}$ to be invertible (thus encoding a change of basis in $\mathbb{R}^n$), we can consider the transformation $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$. A natural question is: what happens to our standard inner product under the transformation? If the standard inner product of $\mathbf{x}, \mathbf{y}$ is $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top\mathbf{y} = \mathbf{x}^\top\mathbf{I}_n\mathbf{y}$, then the standard inner product of the transformed vectors $\mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{y}$ is $\langle \mathbf{A}\mathbf{x}, \mathbf{A}\mathbf{y} \rangle = (\mathbf{A}\mathbf{x})^\top\mathbf{A}\mathbf{y} = \mathbf{x}^\top(\mathbf{A}^\top\mathbf{A})\mathbf{y}$. So $\mathbf{A}^\top\mathbf{A}$ can be thought of as a "scaling factor" by which we can recover the standard inner product in the transformed space under $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$.

Existence of SVD

Let's examine the spectrum[2] of $\mathbf{A}^\top\mathbf{A}$. Since $\mathbf{A}^\top\mathbf{A}$ is PSD, its eigenvalues are all ≥ 0. Since $\mathbf{A}^\top\mathbf{A}$ is symmetric, it has a spectral decomposition

$$\mathbf{A}^\top\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$$

More precisely, if $\mathbf{A}^\top\mathbf{A}$ has r nonzero eigenvalues (with multiplicity), then write

$$\begin{pmatrix} \Lambda' & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{V}' \\ \mathbf{V}_0 \end{pmatrix} \mathbf{A}^\top\mathbf{A} \begin{pmatrix} \mathbf{V}' & \mathbf{V}_0 \end{pmatrix},$$

where Λ' is a diagonal matrix with the nonzero eigenvalues of $\mathbf{A}^\top\mathbf{A}$ corresponding to eigenvectors in $\mathbf{V}'$ and the eigenvectors with vanishing eigenvalue are in $\mathbf{V}_0$.

Next, define

$$\mathbf{U}' = \mathbf{A}\mathbf{V}'\Lambda'^{-\frac{1}{2}} \in \mathbb{R}^{m\times r}.$$

Then we have

$$\begin{aligned} \mathbf{U}'\Lambda'^{\frac{1}{2}}\mathbf{V}'^\top &= \mathbf{A}\mathbf{V}'\Lambda'^{-\frac{1}{2}}\Lambda'^{\frac{1}{2}}\mathbf{V}'^\top \\ &= \mathbf{A}\mathbf{V}'\mathbf{V}'^\top \\ &= \mathbf{A} \text{ because } \mathbf{V} \text{ is unitary.} \end{aligned}$$

Moreover,

$$\begin{aligned} \mathbf{U}'^\top\mathbf{U}' &= (\mathbf{A}\mathbf{V}'\Lambda'^{-\frac{1}{2}})^\top\mathbf{A}\mathbf{V}'\Lambda'^{-\frac{1}{2}} \\ &= \Lambda'^{-\frac{1}{2}}\mathbf{V}'^\top\mathbf{A}^\top\mathbf{A}\mathbf{V}'\Lambda'^{-\frac{1}{2}} \\ &= \Lambda'^{-\frac{1}{2}}\Lambda'\Lambda'^{-\frac{1}{2}} \\ &= \mathbf{I}_r, \end{aligned}$$

so the columns of $\mathbf{U}'$ are orthonormal and can be extended to form an orthonormal basis for $\mathbb{R}^m$. If we choose $\mathbf{U}_0$ containing these added columns, then

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}' & \mathbf{U}_0 \end{pmatrix}$$

is unitary. Next, we form

$$\mathbf{\Sigma} = \begin{pmatrix} \Lambda^{\frac{1}{2}} \\ \mathbf{0} \end{pmatrix}$$

so that $\mathbf{\Sigma}$ has $m - r$ rows of zeros at the bottom and is thus in $\mathbb{R}^{m\times n}$. We arrive at

$$\begin{aligned} \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top &= \begin{pmatrix} \mathbf{U}' & \mathbf{U}_0 \end{pmatrix} \begin{pmatrix} \Lambda^{\frac{1}{2}} \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{V}' \\ \mathbf{V}_0 \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{U}' & \mathbf{U}_0 \end{pmatrix} \begin{pmatrix} \Lambda^{\frac{1}{2}}\mathbf{V}'^\top \\ \mathbf{0} \end{pmatrix} \\ &= \mathbf{U}'\Lambda^{\frac{1}{2}}\mathbf{V}'^\top \\ &= \mathbf{A}. \end{aligned}$$

The decomposition

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

is known as a *singular value decomposition* of $\mathbf{A}$, and we have just proven its existence. To sum up,

[2] A.k.a., the "eigenstuff".

Theorem (SVD)

Given any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, there exist orthonormal bases $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ of $\mathbb{R}^n$ and $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ of $\mathbb{R}^m$ such that if $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$, $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_m)$, and $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ is a rectangular diagonal matrix containing the square roots of the nonzero eigenvalues of $\mathbf{A}^\top \mathbf{A}$, then

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top.$$

The square roots of the nonzero eigenvalues of $\mathbf{A}^\top \mathbf{A}$ are called the *singular values* of $\mathbf{A}$.[3] They are usually denoted (σ_i). The corresponding columns of $\mathbf{V}$, i.e. the eigenvectors $(\mathbf{v}_i)$ of $\mathbf{A}^\top \mathbf{A}$ corresponding to the (σ_i), are called *right-singular vectors* of $\mathbf{A}$. The corresponding columns of $\mathbf{U}$ are called *left-singular vectors* of $\mathbf{A}$. The singular values of $\mathbf{A}$ are unique, but the corresponding singular vectors are not.

Note that the columns of $\mathbf{U}$ form an eigenbasis of $\mathbb{R}^m$ with respect to $\mathbf{A}\mathbf{A}^\top$:

$$\begin{aligned}
\mathbf{A}\mathbf{A}^\top &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top \\
&= \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^\top\mathbf{U}^\top \\
&= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \text{ for appropriately sized } \mathbf{\Lambda}.
\end{aligned}$$

Similarly, we defined $\mathbf{V}$ via the spectral decomposition of $\mathbf{A}^\top \mathbf{A}$:

$$\mathbf{A}^\top \mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^\top \text{ for appropriately sized } \Lambda.$$

From our discussion about $\mathbf{A}^\top \mathbf{A}$, it follows that the number of singular values of $\mathbf{A}$ is equal to $\text{rank}(\mathbf{A})$. The geometric picture for SVD goes as follows:

1. First, via the unitary $\mathbf{V}^\top$, change coordinates to the eigenbasis for $\mathbf{A}^\top \mathbf{A}$.
2. Via $\mathbf{\Sigma}$, which has the same rank as $\mathbf{A}$, scale by the σ_i.
3. Via the unitary $\mathbf{U}$, rotate back.

Note that in Step 1, the actions of the nonsingular $\mathbf{v}_i$ don't matter. Why? Because those $\mathbf{v}_i$ correspond to eigenvalue 0, so they are in $\ker(\mathbf{A}^\top \mathbf{A})$, which we established is the same as $\ker(\mathbf{A})$. More thoroughly,

Theorem (FTLA, Part II)

Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be a singular value decomposition of $\mathbf{A}$, and let $\text{rank}(\mathbf{A}) = r$. Then

(i) The first r columns of $\mathbf{V}$, i.e. the right-singular vectors of $\mathbf{A}$, form an orthonormal basis for $\text{range}(\mathbf{A}^\top)$.

(ii) The last $n - r$ columns of $\mathbf{V}$ form an orthonormal basis for $\ker(\mathbf{A})$.

(iii) The first r columns of $\mathbf{U}$, i.e. the left-singular vectors of $\mathbf{A}$, form an orthonormal basis for $\text{range}(\mathbf{A})$.

(iv) The last $m - r$ columns of $\mathbf{U}$ form an orthonormal basis for $\ker(\mathbf{A}^\top)$.

Proof

Easy to fill in the details from the comment directly preceding the theorem combined with FTLA, Part I. □

When passing to the eigenbasis of $\mathbf{A}^\top \mathbf{A}$ via $\mathbf{V}^\top$, we effectively ignore vectors in $\ker(\mathbf{A})$, as $\mathbf{\Sigma}$ will kill them. As for the relevant coordinates, we let the singular vectors transform them, $\mathbf{\Sigma}$ scale them, and then $\mathbf{U}$ bring them back into the image where they belong.

[3] It's worth noting that $\mathbf{A}^\top \mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$ have the same nonzero eigenvalues.

Application: Principal Component Analysis (PCA)

Motivation

When fitting a model (e.g. an OLS model) to data, we hope to represent the data in a simpler way. A basketball player is a complex object; a vector representing (**points**, **assists**, **rebounds**, **eye color**, **birthday**) is not. However, we hope to use only the most relevant features for (i) fast computation and (ii) to build a more stable model (i.e. reduce the variance of the model). How can we find the features that are important for predicting whether a player helps his team win (probably **points**, **assists**, **rebounds**), allowing us to ignore the features that are not (probably **eye color**, **birthday**)?

Given a matrix of data $\mathbf{X} \in \mathbb{R}^{n \times d}$ containing n-many d-dimensional data points, **PCA** allows us to find a suitable subspace of $\mathbb{R}^d$ onto which we can project our data, leaving us with the most relevant features. How do we decide which features to drop? The idea is that we look at the data and keep only a small number ($< d$) of orthogonal directions (perhaps linear combinations of features) that capture the most variance of the data. Intuitively, the low-variance directions contain less information about the data, so we can throw them away, improving the model's performance on new data without hurting predictive accuracy.

Understanding Variance

The first step of PCA is to center the data so that every feature has mean 0 amongst the data points. We do this because uncentered data would influence our choice of relevant directions (which will be unit-length arrows starting at 0) in unwanted ways. Thus, we first center $\mathbf{X}$ by subtracting the mean vector $\mathbb{E}\mathbf{x} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$ from each row.

Next, we hope to find a unit vector $\mathbf{v} \in \mathbb{R}^d$ that will capture the most "variance in the data." What does this mean? We will ultimately project every data point $\mathbf{x}_i$ onto $\mathbf{v}$ by taking $\mathbf{x}_i^\top \mathbf{v}$, so we care about the variance of this quantity as i ranges through the data from 1 to n.

A brief aside on variance in probabilistic terms: Given a random variable X with $\mathbb{E}X = 0$, we have the following version of Chebyshev's inequality:

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}.$$

This inequality tells us that $\mathrm{Var}(X)$ is a measure of how likely X is to vary from its mean. Variance conveys information about the *tail probabilities* of X; it shows us how likely X is to take a highly unexpected value.

In our case, highly unexpected values correspond to new information. That is why we want $\mathbf{v}$ that maximizes the variance of the $\mathbf{x}_i^\top \mathbf{v}$: it would mean that $\mathrm{span}(\mathbf{v})$ is the 1-dimensional subspace of $\mathbb{R}^d$ containing the most information about the data.

Finding the First Principal Component

We compute the sample variance of this projection amongst our n-many data points:

$$\text{sample variance} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{v})^2 = \frac{1}{n}||\mathbf{X}\mathbf{v}||^2 = \frac{1}{n}\mathbf{v}^\top \mathbf{X}^\top \mathbf{X}\mathbf{v}.$$

Thus, if we want to find the unit vector $\mathbf{v}$ maximizing the variance, we've walked into a constrained optimization problem:

$$\max_{\mathbf{v} \in \mathbb{R}^d} \mathbf{v}^\top \mathbf{X}^\top \mathbf{X}\mathbf{v} \text{ subject to } ||\mathbf{v}|| = 1$$

To solve this optimization problem, recall that for symmetric $\mathbf{A} \in \mathbb{R}^{d \times d}$, for any $\mathbf{v} \in \mathbb{R}^d$ with $||\mathbf{v}|| = 1$,

$$\lambda_{\min}(\mathbf{A}) \leq \mathbf{v}^\top \mathbf{A}\mathbf{v} \leq \lambda_{\max}(\mathbf{A})$$

where for both bounds, equality holds iff $\mathbf{v}$ is a corresponding eigenvector. This immediately yields that the *first loading vector* $\mathbf{v} = \mathbf{v}_1$ is a unit eigenvector corresponding to the maximal eigenvalue of $\mathbf{X}^\top\mathbf{X}$.

Finding More Principal Components

We often want more than one principal component. Given $k-1$ principal components, the problem of finding the k'th amounts to another constrained optimization problem:

$$\max_{\mathbf{v}\in\mathbb{R}^d} \mathbf{v}^\top\mathbf{X}^\top\mathbf{X}\mathbf{v} \text{ subject to } ||\mathbf{v}|| = 1$$
$$\mathbf{v}^\top\mathbf{v}_i = 0 \text{ for } i = 1, \ldots, k-1.$$

The *k'th loading vector* $\mathbf{v}_k$ is given by the following result:

Theorem

The solution to the above optimization problem is $\mathbf{v}$ = a unit eigenvector corresponding to the k'th largest eigenvalue of $\mathbf{X}^\top\mathbf{X}$.

Proof

By induction on k. See source at end of note. □

All this tells us that we can compute the first k loading vectors by computing the SVD of $\mathbf{X}$ and taking the first k right-singular vectors.

Projecting onto the PCA Coordinate System

How do we project the data onto the subspace of $\mathbb{R}^d$ spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_k$? For each data point $\mathbf{x}_i$, we want to map

$$\mathbf{x}_i \mapsto \text{Proj}_{\text{span}\{v_1,\ldots,v_k\}}(\mathbf{x}_i) = \sum_{j=1}^{k}(\mathbf{x}_i^\top\mathbf{v}_j)\mathbf{v}_j.$$

Computationally, it's easier to handle all the $\mathbf{x}_i$ at once through matrix multiplication. Let $\mathbf{V}_k = (\mathbf{v}_1, \ldots, \mathbf{v}_k)$. Then the new data matrix $\widetilde{\mathbf{X}}_k \in \mathbb{R}^{n\times k}$ is given by

$$\widetilde{\mathbf{X}}_k = \mathbf{X}\mathbf{V}_k\mathbf{V}_k^\top.$$

The rows of $\widetilde{\mathbf{X}}_k$ are exactly what we wanted: the original data points projected onto the subspace spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

Geometric View of PCA

OLS could be viewed intuitively as finding the "line of best fit." In it, we minimize the vertical distance between the data points and the fitted line. Similarly, we can view PCA as finding the "subspace of best fit" insofar as the k-dimensional subspace we project onto minimizes perpendicular distance between it and the original data points in $\mathbb{R}^d$.

To show this, we need to show that our first loading vector minimizes the reconstruction error

$$\sum_{i=1}^{n} ||\mathbf{x}_i - P_\mathbf{v}(\mathbf{x}_i)||^2$$

where $P_\mathbf{v}(\mathbf{x}_i) = (\mathbf{x}_i^\top\mathbf{v})\mathbf{v}$ represents the projection of $\mathbf{x}_i$ onto the span of $\mathbf{v}$.

By the Pythagorean theorem, we have that

$$||\mathbf{x}_i - P_\mathbf{v}(\mathbf{x}_i)||^2 + ||P_\mathbf{v}(\mathbf{x}_i)||^2 = ||\mathbf{x}_i||^2$$

so that

$$\sum_{i=1}^{n} ||\mathbf{x}_i - P_{\mathbf{v}}(\mathbf{x}_i)||^2 = \sum_{i=1}^{n} ||\mathbf{x}_i||^2 - ||P_{\mathbf{v}}(\mathbf{x}_i)||^2$$
$$= \sum_{i=1}^{n} ||\mathbf{x}_i||^2 - \sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{v})^2.$$

The first term is constant in $\mathbf{v}$, so minimizing reconstruction error amounts to minimizing $\sum_{i=1}^{n} (\mathbf{x}_i^\top \mathbf{v})^2$, which is precisely our objective in PCA.

Low-Rank Approximation

Let $||\cdot||$ be any unitary-invariant norm on $\mathbb{R}^{n \times d}$. A family of such norms is the collection of induced ℓ^p-norms for matrices:

$$||\mathbf{A}||_p = \sup_{\mathbf{x} \neq 0} \frac{||\mathbf{A}\mathbf{x}||_p}{||\mathbf{x}||_p}$$

of which the operator ($p = 1$) and the spectral ($p = 2$) norms are examples.

Take a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where $\mathbb{R}^{n \times d}$ is equipped with our unitary-invariant norm $|| \cdot ||$. If we seek the best rank-k approximation to $\mathbf{X}$ with respect to $|| \cdot ||$, then PCA comes in handy:

Theorem (Eckart-Young-Mirsky)

Our PCA solution $\widetilde{\mathbf{X}}_k$ is the best rank-k approximation to $\mathbf{X}$ with respect to $|| \cdot ||$ in the sense that for any rank-r ($r \leq k$) matrix $\mathbf{Y} \in \mathbb{R}^{n \times d}$,

$$||\mathbf{X} - \widetilde{\mathbf{X}}_k|| \leq ||\mathbf{X} - \mathbf{Y}||.$$

This theorem tells us that the process of projecting our data onto a subspace via PCA amounts to finding the best rank-k approximation of $\mathbf{X}$.