Background Material

DS-GA 1013 / MATH-GA 2824 Mathematical Tools for Data Science

https://cims.nyu.edu/~cfgranda/pages/MTDS_spring20/index.html

Sreyas Mohan and Carlos Fernandez-Granda

Vector space

Consists of:

- A set $\mathcal{V}$
- A scalar field (usually $\mathbb{R}$ or $\mathbb{C}$)
- Two operations $+$ and $\cdot$

Properties

- For any $\vec{x}, \vec{y} \in \mathcal{V}$, $\vec{x} + \vec{y}$ belongs to $\mathcal{V}$
- For any $\vec{x} \in \mathcal{V}$ and any scalar α, $\alpha \cdot \vec{x} \in \mathcal{V}$
- There exists a zero vector $\vec{0}$ such that $\vec{x} + \vec{0} = \vec{x}$ for any $\vec{x} \in \mathcal{V}$
- For any $\vec{x} \in \mathcal{V}$ there exists an additive inverse $\vec{y}$ such that $\vec{x} + \vec{y} = \vec{0}$, usually denoted by $-\vec{x}$

Properties

- The vector sum is commutative and associative, i.e. for all $\vec{x}, \vec{y}, \vec{z} \in \mathcal{V}$

$$\vec{x} + \vec{y} = \vec{y} + \vec{x}, \quad (\vec{x} + \vec{y}) + \vec{z} = \vec{x} + (\vec{y} + \vec{z})$$

- Scalar multiplication is associative, for any scalars α and β and any $\vec{x} \in \mathcal{V}$

$$\alpha\,(\beta \cdot \vec{x}) = (\alpha\,\beta) \cdot \vec{x}$$

- Scalar and vector sums are both distributive, i.e. for any scalars α and β and any $\vec{x}, \vec{y} \in \mathcal{V}$

$$(\alpha + \beta) \cdot \vec{x} = \alpha \cdot \vec{x} + \beta \cdot \vec{x}, \quad \alpha \cdot (\vec{x} + \vec{y}) = \alpha \cdot \vec{x} + \alpha \cdot \vec{y}$$

Concept Check

Let $\mathcal{V} = \{x | x \in \mathbb{R}, x \geq 0\}$. Define addition operation for $x, y \in \mathcal{V}$ as $x + y = x + y$ (normal addition) and scalar multiplication for $x \in \mathcal{V}$ and $\alpha \in \mathbb{R}$ as $\alpha x = \alpha . x$ (regular scaling). Is $\mathcal{V}$ a vector field?

Subspaces

A subspace of a vector space $\mathcal{V}$ is any subset of $\mathcal{V}$ that is *also itself a vector space*

Linear dependence/independence

A set of m vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_m$ is linearly dependent if there exist m scalar coefficients $\alpha_1, \alpha_2, \ldots, \alpha_m$ which are not all equal to zero and

$$\sum_{i=1}^{m} \alpha_i \, \vec{x}_i = \vec{0}$$

Equivalently, any vector in a linearly dependent set can be expressed as a linear combination of the rest

Span

The span of $\{\vec{x}_1, \ldots, \vec{x}_m\}$ is the set of all possible linear combinations

$$\text{span}(\vec{x}_1, \ldots, \vec{x}_m) := \left\{ \vec{y} \mid \vec{y} = \sum_{i=1}^{m} \alpha_i \, \vec{x}_i \quad \text{for some scalars } \alpha_1, \alpha_2, \ldots, \alpha_m \right\}$$

The span of any set of vectors in $\mathcal{V}$ is a subspace of $\mathcal{V}$

Basis and dimension

A basis of a vector space $\mathcal{V}$ is a set of independent vectors $\{\vec{x}_1, \ldots, \vec{x}_m\}$ such that

$$\mathcal{V} = \text{span}\left(\vec{x}_1, \ldots, \vec{x}_m\right)$$

If $\mathcal{V}$ has a basis with finite cardinality then every basis contains the same number of vectors

The dimension $\dim\left(\mathcal{V}\right)$ of $\mathcal{V}$ is the cardinality of any of its bases

Equivalently, the dimension is the number of linearly independent vectors that span $\mathcal{V}$

Standard basis

$$\vec{e}_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \vec{e}_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \quad \ldots, \quad \vec{e}_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}$$

The dimension of $\mathbb{R}^n$ is n

Concept Check

- (True/False) If S is a subset of vector space $\mathcal{V}$, then $span(S)$ contains the intersection of all subspace of $\mathcal{V}$ that contain S.
- The set of all $n \times n$ matrices with trace as zero forms a subspace W of the space of $n \times n$ matrices. Find a basis for W and calculate it's dimension.

Concept Check - Answers

- True.
- We need to enforce that the sum of diagonal entries is zero, or that $A_{11} + A_{22} + \cdots + A_{nn} = 0$. The basis vectors can be $\{E_{ij}\}_{i \neq j} \cup \{E_{ii} - E_{nn}\}_{i=1,2,\ldots,n-1}$. The dimension of W is $n^2 - 1$

Inner product

Operation $\langle \cdot, \cdot \rangle$ that maps a pair of vectors to a scalar

Properties

- If the scalar field is $\mathbb{R}$, it is symmetric. For any $\vec{x}, \vec{y} \in \mathcal{V}$

  $$\langle \vec{x}, \vec{y} \rangle = \langle \vec{y}, \vec{x} \rangle$$

  If the scalar field is $\mathbb{C}$, then for any $\vec{x}, \vec{y} \in \mathcal{V}$

  $$\langle \vec{x}, \vec{y} \rangle = \overline{\langle \vec{y}, \vec{x} \rangle},$$

  where for any $\alpha \in \mathbb{C}$ $\overline{\alpha}$ is the complex conjugate of α

Properties

- It is linear in the first argument, i.e. for any $\alpha \in \mathbb{R}$ and any $\vec{x}, \vec{y}, \vec{z} \in \mathcal{V}$

$$\langle \alpha \, \vec{x}, \vec{y} \rangle = \alpha \, \langle \vec{x}, \vec{y} \rangle \,,$$
$$\langle \vec{x} + \vec{y}, \vec{z} \rangle = \langle \vec{x}, \vec{z} \rangle + \langle \vec{y}, \vec{z} \rangle \,.$$

  If the scalar field is $\mathbb{R}$, it is also linear in the second argument

- It is positive definite: $\langle \vec{x}, \vec{x} \rangle$ is nonnegative for all $\vec{x} \in \mathcal{V}$ and if $\langle \vec{x}, \vec{x} \rangle = 0$ then $\vec{x} = \vec{0}$

Dot product

Inner product between $\vec{x}, \vec{y} \in \mathbb{R}^n$

$$\vec{x} \cdot \vec{y} := \sum_i \vec{x}[i] \; \vec{y}[i]$$

$\mathbb{R}^n$ endowed with the dot product is usually called a Euclidean space of dimension n

If $\vec{x}, \vec{y} \in \mathbb{C}^n$

$$\vec{x} \cdot \vec{y} := \sum_i \vec{x}[i] \; \overline{\vec{y}[i]}$$

Matrix inner product

The inner product between two $m \times n$ matrices A and B is

$$\langle A, B \rangle := \operatorname{tr}\left(A^T B\right) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij}$$

where the trace of an $n \times n$ matrix is defined as the sum of its diagonal

$$\operatorname{tr}(M) := \sum_{i=1}^{n} M_{ii}$$

For any pair of $m \times n$ matrices A and B

$$\operatorname{tr}\left(B^T A\right) := \operatorname{tr}\left(A B^T\right)$$

Function inner product

The inner product between two complex-valued square-integrable functions f, g defined in an interval $[a, b]$ of the real line is

$$\vec{f} \cdot \vec{g} := \int_a^b f(x)\,\overline{g(x)}\,\mathrm{d}x$$

Norms

Let $\mathcal{V}$ be a vector space, a norm is a function $||\cdot||$ from $\mathcal{V}$ to $\mathbb{R}$ with the following properties

- It is homogeneous. For any scalar α and any $\vec{x} \in \mathcal{V}$

$$||\alpha \, \vec{x}|| = |\alpha| \, ||\vec{x}|| \, .$$

- It satisfies the triangle inequality

$$||\vec{x} + \vec{y}|| \leq ||\vec{x}|| + ||\vec{y}|| \, .$$

  In particular, $||\vec{x}|| \geq 0$

- $||\vec{x}|| = 0$ implies $\vec{x} = \vec{0}$

Inner-product norm

Square root of inner product of vector with itself

$$||\vec{x}||_{\langle\cdot,\cdot\rangle} := \sqrt{\langle\vec{x},\vec{x}\rangle}$$

Inner-product norm

- Vectors in $\mathbb{R}^n$ or $\mathbb{C}^n$: ℓ_2 norm

$$||\vec{x}||_2 := \sqrt{\vec{x} \cdot \vec{x}} = \sqrt{\sum_{i=1}^{n} \vec{x}[i]^2}$$

- Matrices in $\mathbb{R}^{m \times n}$ or $\mathbb{C}^{m \times n}$: Frobenius norm

$$||A||_{\text{F}} := \sqrt{\text{tr}\left(A^T A\right)} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2}$$

- Square-integrable complex-valued functions: $\mathcal{L}_2$ norm

$$||f||_{\mathcal{L}_2} := \sqrt{\langle f, f \rangle} = \sqrt{\int_a^b |f(x)|^2 \, dx}$$

Cauchy-Schwarz inequality

For any two vectors $\vec{x}$ and $\vec{y}$ in an inner-product space

$$|\langle \vec{x}, \vec{y} \rangle| \leq ||\vec{x}||_{\langle \cdot, \cdot \rangle} \, ||\vec{y}||_{\langle \cdot, \cdot \rangle}$$

Assume $||\vec{x}||_{\langle \cdot, \cdot \rangle} \neq 0$, then

$$\langle \vec{x}, \vec{y} \rangle = - ||\vec{x}||_{\langle \cdot, \cdot \rangle} \, ||\vec{y}||_{\langle \cdot, \cdot \rangle} \iff \vec{y} = - \frac{||\vec{y}||_{\langle \cdot, \cdot \rangle}}{||\vec{x}||_{\langle \cdot, \cdot \rangle}} \vec{x}$$

$$\langle \vec{x}, \vec{y} \rangle = ||\vec{x}||_{\langle \cdot, \cdot \rangle} \, ||\vec{y}||_{\langle \cdot, \cdot \rangle} \iff \vec{y} = \frac{||\vec{y}||_{\langle \cdot, \cdot \rangle}}{||\vec{x}||_{\langle \cdot, \cdot \rangle}} \vec{x}$$

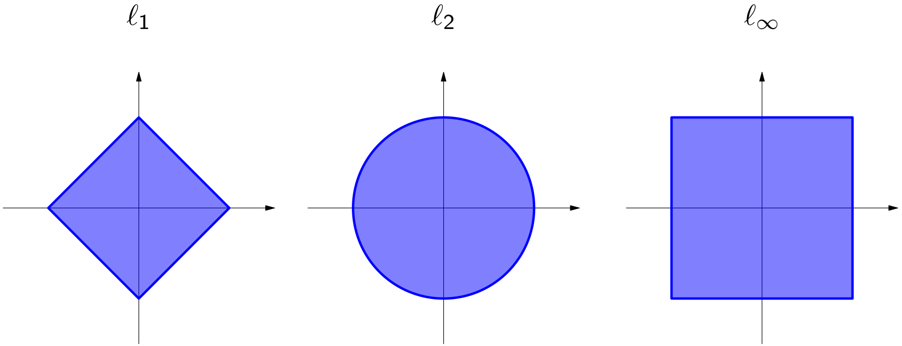
ℓ_1 and ℓ_∞ norms

Norms in $\mathbb{R}^n$ or $\mathbb{C}^n$ not induced by an inner product

$$||\vec{x}||_1 := \sum_{i=1}^{n} |\vec{x}[i]|$$

$$||\vec{x}||_\infty := \max_i |\vec{x}[i]|$$

Norm balls



Distance

The distance between two vectors $\vec{x}$ and $\vec{y}$ induced by a norm $||\cdot||$ is

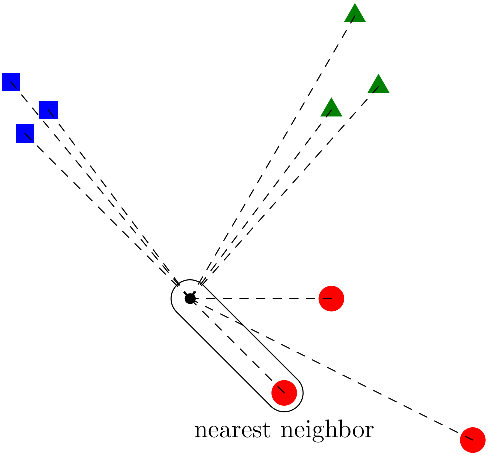
$$d\left(\vec{x}, \vec{y}\right) := ||\vec{x} - \vec{y}||$$

Classification

Aim: Assign a signal to one of k predefined classes

Training data: n pairs of signals (represented as vectors) and labels: $\{\vec{x}_1, l_1\}, \ldots, \{\vec{x}_n, l_n\}$

Nearest-neighbor classification



Face recognition

Training set: 360 64×64 images from 40 different subjects (9 each)

Test set: 1 new image from each subject

We model each image as a vector in $\mathbb{R}^{4096}$ and use the ℓ_2-norm distance

Face recognition

Training set

Nearest-neighbor classification

Errors: 4 / 40

Test image

Closest image

Mean, Variance and Correlation

- Consider real-valued data corresponding to a single quantity or feature. We model such data as a scalar continuous random variable.
- In reality we usually have access to a finite number of data points, not to a continuous distribution.
- Mean of a random variable is the point that minimizes the expected distance to the random variable.
- Intuitively, it is the center of mass of the probability density, and hence of the dataset.

Mean

Lemma: For any random variable $\tilde{a}$ with mean $\mathrm{E}(\tilde{a})$,

$$\mathrm{E}\left(\tilde{a}\right) = \arg\min_{c \in \mathbb{R}} \mathrm{E}\left((c - \tilde{a})^2\right).$$

Proof

Let $g(c) := \mathrm{E}\left((c - \tilde{a})^2\right) = c^2 - 2c\mathrm{E}\left(\tilde{a}\right) + \mathrm{E}\left(\tilde{a}^2\right)$, we have

$$f'(c) = 2(c - \mathrm{E}(\tilde{a})),$$
$$f''(c) = 2.$$

The function is strictly convex and has a minimum where the derivative equals zero, i.e. when c is equal to the mean.

Variance

The variance of a random variable $\tilde{a}$

$$\mathrm{Var}(\tilde{a}) := \mathrm{E}\left((\tilde{a} - \mathrm{E}(\tilde{a}))^2\right)$$

quantifies how much it fluctuates around its mean. The standard deviation, defined as the square root of the variance, is therefore a measure of how spread out the dataset is around its center.

Covariance

- Consider data containing two features, each represented by a random variable.
- The covariance of two random variables $\tilde{a}$ and $\tilde{b}$ quantifies their joint fluctuations around their respective means.

$$\mathrm{Cov}(\tilde{a}, \tilde{b}) := \mathrm{E}\left[(\tilde{a} - \mathrm{E}(\tilde{a})(\tilde{b} - \mathrm{E}(\tilde{b}))\right]$$

Concept Check: Zero Mean RVs

- The space of zero mean random variables form a vector space. Why?
- What will be the origin (zero vector) of the space?
- Does $\mathrm{Cov}(\tilde{a}, \tilde{b})$ define a valid inner product in this space?

Vector Space of Zero Mean RVs

- Zero-mean random variables form a vector space because linear combinations of zero-mean random variables are also zero mean.
- The origin of the vector space (the zero vector) is the random variable equal to zero with probability one.
- The covariance is a valid inner product because it is (1) symmetric, (2) linear in its first argument, i.e. for any $\alpha \in \mathbb{R}$ $\mathrm{E}(\alpha \tilde{a}\tilde{b}) = \alpha \mathrm{E}(\tilde{a}\tilde{b})$, and (3) positive definite, i.e. $\mathrm{E}(\tilde{a}^2) > 0$ if $\tilde{a} \neq 0$ and $\mathrm{E}(\tilde{a}^2) = 0$ if and only if $\tilde{a} = 0$ with probability one. To prove this last property, we use a fundamental inequality in probability theory.

Markov's Inequality

Theorem (Markov's inequality)

Let $\tilde{r}$ be a nonnegative random variable. For any positive constant $c > 0$,

$$\mathrm{P}(\tilde{r} \geq c) \leq \frac{\mathrm{E}(\tilde{r})}{c}.$$

Proof

Consider the indicator variable $1_{\tilde{r} \geq c}$. We have

$$\tilde{r} - c\,1_{\tilde{r} \geq c} \geq 0,$$

Proof

Consider the indicator variable $1_{\tilde{r} \geq c}$. We have

$$\tilde{r} - c\, 1_{\tilde{r} \geq c} \geq 0,$$

By linearity of expectation and the fact that $1_{\tilde{r} \geq c}$ is a Bernoulli random variable with expectation $\mathrm{P}(\tilde{r} \geq c)$ we have

$$\mathrm{E}(\tilde{r}) \geq c\, \mathrm{E}\left(1_{\tilde{r} \geq c}\right) = c\, \mathrm{P}(\tilde{r} \geq c).$$

Corollary

If the mean square $\mathrm{E}\left[\tilde{a}^2\right]$ of a random variable $\tilde{a}$ equals zero, then

$$\mathrm{P}(\tilde{a} \neq 0) = 0.$$

Corollary

If the mean square $\mathrm{E}\left[\tilde{a}^2\right]$ of a random variable $\tilde{a}$ equals zero, then

$$\mathrm{P}(\tilde{a} \neq 0) = 0.$$

Proof:

- If $\mathrm{P}(\tilde{a} \neq 0) \neq 0$ then there exists an ϵ such that $\mathrm{P}(\tilde{a}^2 \geq \epsilon) \neq 0$.

Corollary

If the mean square $\mathrm{E}\left[\tilde{a}^2\right]$ of a random variable $\tilde{a}$ equals zero, then

$$\mathrm{P}(\tilde{a} \neq 0) = 0.$$

Proof:

- If $\mathrm{P}(\tilde{a} \neq 0) \neq 0$ then there exists an ϵ such that $\mathrm{P}(\tilde{a}^2 \geq \epsilon) \neq 0$.
- This is impossible.
- Applying Markov's inequality to the nonnegative random variable $\tilde{a}^2$ we have

$$\begin{aligned}\mathrm{P}(\tilde{a}^2 \geq \epsilon) &\leq \frac{\mathrm{E}\left(\tilde{a}^2\right)}{\epsilon} \\ &= 0.\end{aligned}$$

Correlation Coefficient

- When comparing two vectors, a natural measure of their similarity is the cosine of the angle between them which ranges from -1 to 1.
- The cosine equals the inner product between the vectors normalized by their norms.
- In the vector space of zero-mean random variables this quantity is called the correlation coefficient,

$$\rho_{\tilde{a},\tilde{b}} := \frac{\mathrm{Cov}(\tilde{a},\tilde{b})}{\sqrt{\mathrm{Var}(\tilde{a})\mathrm{Var}(\tilde{b})}},$$

Correlation Coefficient

- When comparing two vectors, a natural measure of their similarity is the cosine of the angle between them which ranges from -1 to 1.
- The cosine equals the inner product between the vectors normalized by their norms.
- In the vector space of zero-mean random variables this quantity is called the correlation coefficient,

$$\rho_{\tilde{a},\tilde{b}} := \frac{\mathrm{Cov}(\tilde{a},\tilde{b})}{\sqrt{\mathrm{Var}(\tilde{a})\mathrm{Var}(\tilde{b})}},$$

- $-1 \leq \rho_{\tilde{a},\tilde{b}} \leq 1$. Why?

Cauchy-Schwarz inequality for random variables

Theorem (Cauchy-Schwarz inequality for random variables)

Let $\tilde{a}$ and $\tilde{b}$ be two random variables. Their correlation coefficient satisfies

$$-1 \leq \rho_{\tilde{a},\tilde{b}} \leq 1$$

with equality if and only if $\tilde{b}$ is a linear function of $\tilde{a}$ with probability one.

Proof

Consider the standardized random variables (centered and normalized),

$$\mathsf{s}(\tilde{a}) := \frac{\tilde{a} - \mathrm{E}(\tilde{a})}{\sqrt{\mathrm{Var}(\tilde{a})}}, \qquad \mathsf{s}(\tilde{b}) := \frac{\tilde{b} - \mathrm{E}(\tilde{b})}{\sqrt{\mathrm{Var}(\tilde{b})}}.$$

Proof

Consider the standardized random variables (centered and normalized),

$$\mathsf{s}(\tilde{a}) := \frac{\tilde{a} - \mathrm{E}(\tilde{a})}{\sqrt{\mathrm{Var}(\tilde{a})}}, \qquad \mathsf{s}(\tilde{b}) := \frac{\tilde{b} - \mathrm{E}(\tilde{b})}{\sqrt{\mathrm{Var}(\tilde{b})}}.$$

The mean square distance between them equals

$$\begin{aligned} \mathrm{E}\left[(\mathsf{s}(\tilde{b}) - \mathsf{s}(\tilde{a}))^2\right] &= \mathrm{E}\left(\mathsf{s}(\tilde{a})^2\right) + \mathrm{E}(\mathsf{s}(\tilde{b})^2) - 2\mathrm{E}(\mathsf{s}(\tilde{a})\,\mathsf{s}(\tilde{b})) \\ &= 2(1 - \mathrm{E}(\mathsf{s}(\tilde{a})\,\mathsf{s}(\tilde{b}))) \\ &= 2(1 - \rho_{\tilde{a},\tilde{b}}) \end{aligned}$$

This implies that $\rho_{\tilde{a},\tilde{b}} \leq 1$. Why?

Proof

- $$\mathrm{E}\left[(\mathsf{s}(\tilde{b}) - \mathsf{s}(\tilde{a}))^2\right] = 2(1 - \rho_{\tilde{a},\tilde{b}})$$
- Recall that if the mean square $\mathrm{E}\left[\tilde{a}^2\right]$ of a random variable $\tilde{a}$ equals zero, then $\mathrm{P}(\tilde{a} \neq 0) = 0$.
- When $\rho_{\tilde{a},\tilde{b}} = 1$, $\mathrm{E}\left[(\mathsf{s}(\tilde{b}) - \mathsf{s}(\tilde{a}))^2\right] = 0$. This means that $\mathsf{s}(\tilde{a}) = \mathsf{s}(\tilde{b})$ with probability one, which implies the linear relationship.

Proof

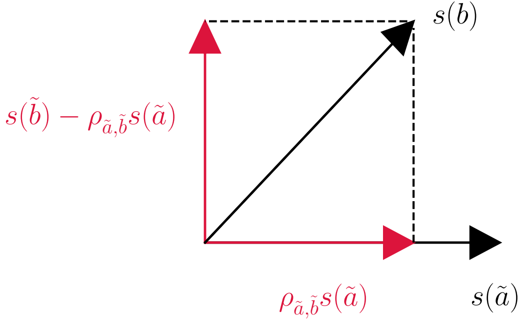
- $$\mathrm{E}\left[(\mathsf{s}(\tilde{b}) - \mathsf{s}(\tilde{a}))^2\right] = 2(1 - \rho_{\tilde{a},\tilde{b}})$$

- Recall that if the mean square $\mathrm{E}\left[\tilde{a}^2\right]$ of a random variable $\tilde{a}$ equals zero, then $\mathrm{P}(\tilde{a} \neq 0) = 0$.
- When $\rho_{\tilde{a},\tilde{b}} = 1$, $\mathrm{E}\left[(\mathsf{s}(\tilde{b}) - \mathsf{s}(\tilde{a}))^2\right] = 0$. This means that $\mathsf{s}(\tilde{a}) = \mathsf{s}(\tilde{b})$ with probability one, which implies the linear relationship.
- Similarly, using

$$\mathrm{E}\left[(\mathsf{s}(\tilde{b}) - (-\,\mathsf{s}(\tilde{a})))^2\right] = 2(1 + \rho_{\tilde{a},\tilde{b}}).$$

  the same argument applies when $\rho_{\tilde{a},\tilde{b}} = -1$.

Geometric Interpretation of Correlation Coefficient



Sample mean, variance and correlation

- When analyzing data we do not have access to a probability distribution, but rather to a set of points.
- Adapt our previous analysis to this setting.
- **Main Idea**: Approximate expectations by averaging over the data

Sample mean, variance and correlation

- Consider a dataset containing n real-valued data with two real valued features $(a_1, b_1), \ldots, (a_n, b_n)$. Let $\mathcal{A} := \{a_1, \ldots, a_n\}$ and $\mathcal{B} := \{b_1, \ldots, b_n\}$

- Sample Mean:

$$\text{av}\left(\mathcal{A}\right) := \frac{1}{n}\sum_{i=1}^{n} a_i,$$

- Sample Covariance

$$\text{cov}(\mathcal{A}, \mathcal{B}) := \frac{1}{n}\sum_{i=1}^{n}(a_i - \text{av}(\mathcal{A}))(b_i - \text{av}(\mathcal{B}),$$

- Sample Variance,

$$\text{var}\left(\mathcal{A}\right) := \frac{1}{n}\sum_{i=1}^{n}\left(a_i - \text{av}\left(\mathcal{A}\right)\right)^2.$$

Sample mean converges to true mean

Theorem (Sample mean converges to true mean)

Let $\tilde{\mathcal{A}}_n$ contain n iid copies $\tilde{a}_1, \ldots, \tilde{a}_n$ of a random variable $\tilde{a}$ with finite variance. Then,

$$\lim_{n} \mathrm{E}\left((\mathrm{av}(\tilde{\mathcal{A}}_n) - \mathrm{E}(\tilde{a}))^2 \right) = 0.$$

Proof

By linearity of expection

$$\mathrm{E}\left(\mathsf{av}(\tilde{\mathcal{A}}_n)\right) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(\tilde{a}_i)$$
$$= \mathrm{E}(\tilde{a}),$$

Proof

By linearity of expection

$$\mathrm{E}\left(\mathsf{av}(\tilde{\mathcal{A}}_n)\right) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\tilde{a}_i)$$
$$= \mathrm{E}(\tilde{a}),$$

which implies

$$\mathrm{E}\left((\mathsf{av}(\tilde{\mathcal{A}}_n) - \mathrm{E}(\tilde{a}))^2\right) = \mathrm{Var}\left(\mathsf{av}(\tilde{\mathcal{A}}_n)\right)$$
$$= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(\tilde{a}_i) \quad \text{by independence}$$
$$= \frac{\mathrm{Var}(\tilde{a})}{n}.$$

The same proof can be applied to the sample variance and the sample covariance, under the assumption that higher-order moments of the distribution are bounded.

Sample Mean is the Center

Lemma (The sample mean is the center)

For any set of real numbers $\mathcal{A} := \{a_1, \ldots, a_n\}$,

$$\operatorname{av}(\mathcal{A}) = \arg\min_{c \in \mathbb{R}} \sum_{i=1}^{n} (c - a_i)^2.$$

Proof

Let $f(c) := \sum_{i=1}^{n}(c - a_i)^2$, we have

$$\begin{aligned} f'(c) &= 2\sum_{i=1}^{n}(c - a_i) \\ &= 2\left(nc - \sum_{i=1}^{n} a_i\right), \\ f''(c) &= 2n. \end{aligned}$$

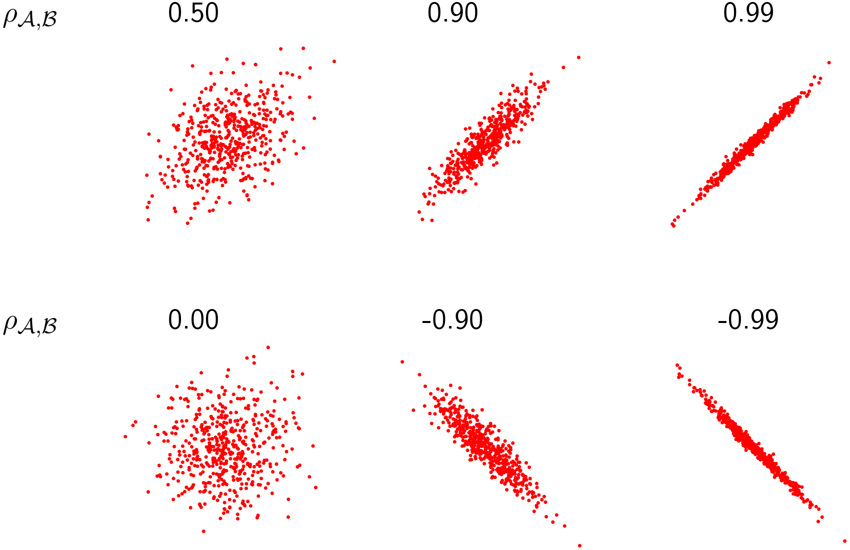
The function is strictly convex and has a minimum where the derivative equals zero, i.e. when c is equal to the sample mean.

Proof

- Note that the proof is essentially the same as that of the probabilistic setting.
- The reason is that both expectation and averaging operators are linear.
- Analogously to the probabilistic setting, we can show that the sample covariance is a valid inner product between centered sets of samples, and the sample standard deviation– defined as the square root of the sample variance– is its corresponding norm.

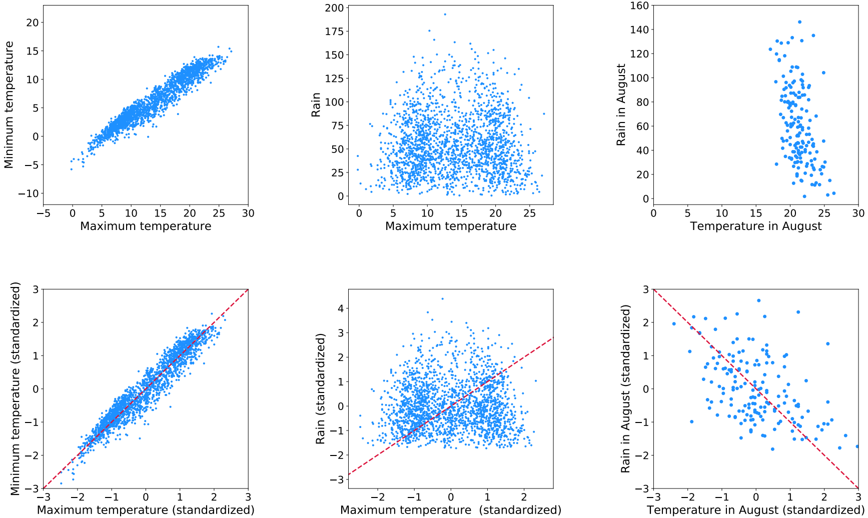
$$\rho_{\mathcal{A},\mathcal{B}} := \frac{\operatorname{cov}(\mathcal{A},\mathcal{B})}{\sqrt{\operatorname{var}(\mathcal{A})\operatorname{var}(\mathcal{B})}}$$

Correlation coefficient



Oxford Data

$\rho = 0.962$ $\rho = 0.019$ $\rho = -0.468$



Oxford Data - Takeaways

- The maximum temperature is highly correlated with the minimum temperature ($\rho = 0.962$).
- Rainfall is almost uncorrelated with the maximum temperature ($\rho = 0.019$), but this **does not mean that the two quantities are not related**; the relation is just not linear.
- When we only consider the rain and temperature in August, then the two quantities are linearly related to some extent. Their correlation is negative ($\rho = -0.468$): when it is warmer it tends to rain less.
- If the relationship between each pair of features were perfectly linearly then they would lie on the dashed red diagonal lines.

Orthogonality

Two vectors $\vec{x}$ and $\vec{y}$ are orthogonal if and only if

$$\langle \vec{x}, \vec{y} \rangle = 0$$

A vector $\vec{x}$ is orthogonal to a set $\mathcal{S}$, if

$$\langle \vec{x}, \vec{s} \rangle = 0, \quad \text{for all } \vec{s} \in \mathcal{S}$$

Two sets of $\mathcal{S}_1, \mathcal{S}_2$ are orthogonal if for any $\vec{x} \in \mathcal{S}_1, \vec{y} \in \mathcal{S}_2$

$$\langle \vec{x}, \vec{y} \rangle = 0$$

The orthogonal complement of a subspace $\mathcal{S}$ is

$$\mathcal{S}^\perp := \{\vec{x} \mid \langle \vec{x}, \vec{y} \rangle = 0 \quad \text{for all } \vec{y} \in \mathcal{S}\}$$

Pythagorean theorem

If $\vec{x}$ and $\vec{y}$ are orthogonal

$$||\vec{x}+\vec{y}||^2_{\langle\cdot,\cdot\rangle} = ||\vec{x}||^2_{\langle\cdot,\cdot\rangle} + ||\vec{y}||^2_{\langle\cdot,\cdot\rangle}$$

Orthonormal basis

Basis of mutually orthogonal vectors with inner-product norm equal to one

If $\{\vec{u}_1, \ldots, \vec{u}_n\}$ is an orthonormal basis of a vector space $\mathcal{V}$, for any $\vec{x} \in \mathcal{V}$

$$\vec{x} = \sum_{i=1}^{n} \langle \vec{u}_i, \vec{x} \rangle \, \vec{u}_i$$

Gram-Schmidt

Builds orthonormal basis from a set of linearly independent vectors $\vec{x}_1, \ldots, \vec{x}_m$ in $\mathbb{R}^n$

1. Set $\vec{u}_1 := \vec{x}_1 / \left|\left|\vec{x}_1\right|\right|_2$

2. For $i = 1, \ldots, m$, compute

$$\vec{v}_i := \vec{x}_i - \sum_{j=1}^{i-1} \langle \vec{u}_j, \vec{x}_i \rangle \, \vec{u}_j$$

   and set $\vec{u}_i := \vec{v}_i / \left|\left|\vec{v}_i\right|\right|_2$

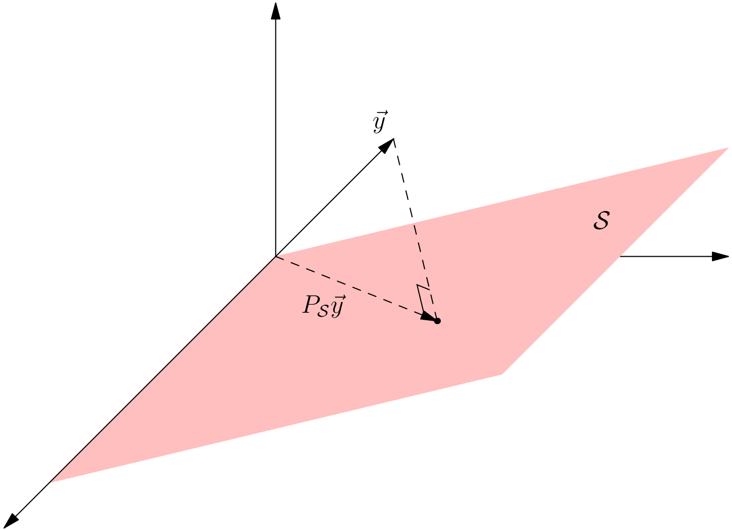
Orthogonal projection

The orthogonal projection of $\vec{x}$ onto a subspace $\mathcal{S}$ is a vector denoted by $\mathcal{P}_{\mathcal{S}}\,\vec{x}$ such that

$$\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{x} \in \mathcal{S}^{\perp}$$

The orthogonal projection is unique

Orthogonal projection



Orthogonal projection

Any vector $\vec{x}$ can be decomposed into

$$\vec{x} = \mathcal{P}_{\mathcal{S}}\,\vec{x} + \mathcal{P}_{\mathcal{S}^{\perp}}\,\vec{x}.$$

For any orthonormal basis $\vec{b}_1, \ldots, \vec{b}_m$ of $\mathcal{S}$,

$$\mathcal{P}_{\mathcal{S}}\,\vec{x} = \sum_{i=1}^{m} \left\langle \vec{x}, \vec{b}_i \right\rangle \vec{b}_i$$

The orthogonal projection is a linear operation. For $\vec{x}$ and $\vec{y}$

$$\mathcal{P}_{\mathcal{S}}\,(\vec{x} + \vec{y}) = \mathcal{P}_{\mathcal{S}}\,\vec{x} + \mathcal{P}_{\mathcal{S}}\,\vec{y}$$

Orthogonal projection is closest

The orthogonal projection $\mathcal{P}_{\mathcal{S}}\,\vec{x}$ of a vector $\vec{x}$ onto a subspace $\mathcal{S}$ is the solution to the optimization problem

$$\begin{aligned} \underset{\vec{u}}{\text{minimize}} \qquad & ||\vec{x} - \vec{u}||_{\langle\cdot,\cdot\rangle} \\ \text{subject to} \qquad & \vec{u} \in \mathcal{S} \end{aligned}$$

Proof

Take any point $\vec{s} \in \mathcal{S}$ such that $\vec{s} \neq \mathcal{P}_{\mathcal{S}}\, \vec{x}$

$$||\vec{x} - \vec{s}||^2_{\langle \cdot, \cdot \rangle}$$

Proof

Take any point $\vec{s} \in \mathcal{S}$ such that $\vec{s} \neq \mathcal{P}_{\mathcal{S}}\, \vec{x}$

$$||\vec{x} - \vec{s}||^2_{\langle \cdot, \cdot \rangle} = ||\vec{x} - \mathcal{P}_{\mathcal{S}}\, \vec{x} + \mathcal{P}_{\mathcal{S}}\, \vec{x} - \vec{s}||^2_{\langle \cdot, \cdot \rangle}$$

Proof

Take any point $\vec{s} \in \mathcal{S}$ such that $\vec{s} \neq \mathcal{P}_{\mathcal{S}}\,\vec{x}$

$$
\begin{aligned}
||\vec{x} - \vec{s}||^2_{\langle\cdot,\cdot\rangle} &= ||\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{x} + \mathcal{P}_{\mathcal{S}}\,\vec{x} - \vec{s}||^2_{\langle\cdot,\cdot\rangle} \\
&= ||\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{x}||^2_{\langle\cdot,\cdot\rangle} + ||\mathcal{P}_{\mathcal{S}}\,\vec{x} - \vec{s}||^2_{\langle\cdot,\cdot\rangle}
\end{aligned}
$$

Proof

Take any point $\vec{s} \in \mathcal{S}$ such that $\vec{s} \neq \mathcal{P}_{\mathcal{S}}\,\vec{x}$

$$
\begin{aligned}
||\vec{x} - \vec{s}||^2_{\langle\cdot,\cdot\rangle} &= ||\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{x} + \mathcal{P}_{\mathcal{S}}\,\vec{x} - \vec{s}||^2_{\langle\cdot,\cdot\rangle} \\
&= ||\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{x}||^2_{\langle\cdot,\cdot\rangle} + ||\mathcal{P}_{\mathcal{S}}\,\vec{x} - \vec{s}||^2_{\langle\cdot,\cdot\rangle} \\
&> ||\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{x}||^2_{\langle\cdot,\cdot\rangle} \quad \text{if } \vec{s} \neq \mathcal{P}_{\mathcal{S}}\,\vec{x}
\end{aligned}
$$

Denoising

Aim: Estimating a signal from perturbed measurements

If the noise is additive, the data are modeled as the sum of the signal $\vec{x}$ and a perturbation $\vec{z}$

$$\vec{y} := \vec{x} + \vec{z}$$

The goal is to estimate $\vec{x}$ from $\vec{y}$

Assumptions about the signal and noise structure are necessary

Denoising via orthogonal projection

Assumption: Signal is well approximated as belonging to a predefined subspace $\mathcal{S}$

Estimate: $\mathcal{P}_{\mathcal{S}}\,\vec{y}$, orthogonal projection of the noisy data onto $\mathcal{S}$

Error:

$$||\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{y}||_2^2 = ||\mathcal{P}_{\mathcal{S}^\perp}\,\vec{x}||_2^2 + ||\mathcal{P}_{\mathcal{S}}\,\vec{z}||_2^2$$

Proof

$$\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{y}$$

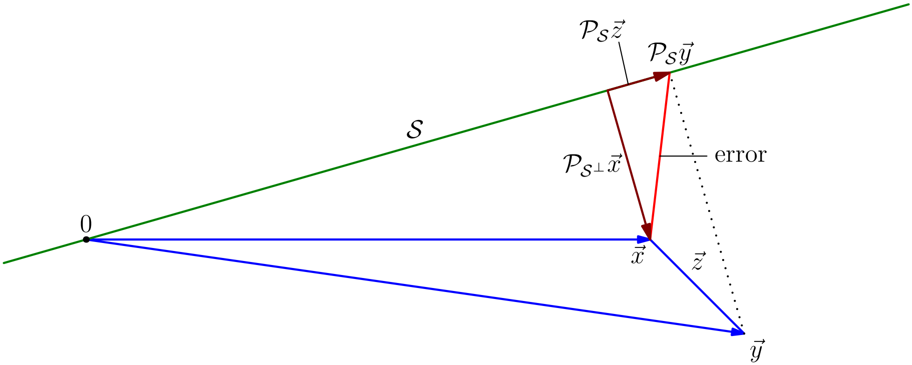
Proof

$$\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{y} = \vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{z}$$

Proof

$$\begin{aligned}\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{y} &= \vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{z} \\ &= \mathcal{P}_{\mathcal{S}^{\perp}}\,\vec{x} - \mathcal{P}_{\mathcal{S}}\,\vec{z}\end{aligned}$$

Error



$\mathcal{P}_{\mathcal{S}}\vec{z}$
$\mathcal{P}_{\mathcal{S}}\vec{y}$
$\mathcal{S}$
$\mathcal{P}_{\mathcal{S}^\perp}\vec{x}$
error
0
$\vec{x}$
$\vec{z}$
$\vec{y}$

Face denoising

Training set: 360 64×64 images from 40 different subjects (9 each)

Noise: iid Gaussian noise

$$\text{SNR} := \frac{||\vec{x}||_2}{||\vec{z}||_2} = 6.67$$

We model each image as a vector in $\mathbb{R}^{4096}$

Face denoising

We denoise by projecting onto:

- $\mathcal{S}_1$: the span of the 9 images from the same subject
- $\mathcal{S}_2$: the span of the 360 images in the training set

Test error:

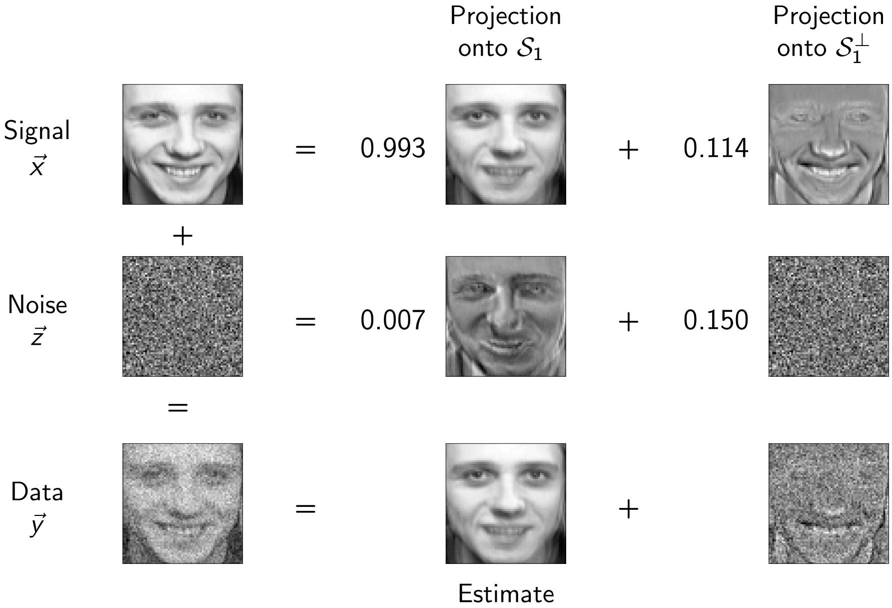
$$\frac{||\vec{x} - \mathcal{P}_{\mathcal{S}_1} \vec{y}||_2}{||\vec{x}||_2} = 0.114$$

$$\frac{||\vec{x} - \mathcal{P}_{\mathcal{S}_2} \vec{y}||_2}{||\vec{x}||_2} = 0.078$$

$\mathcal{S}_1$

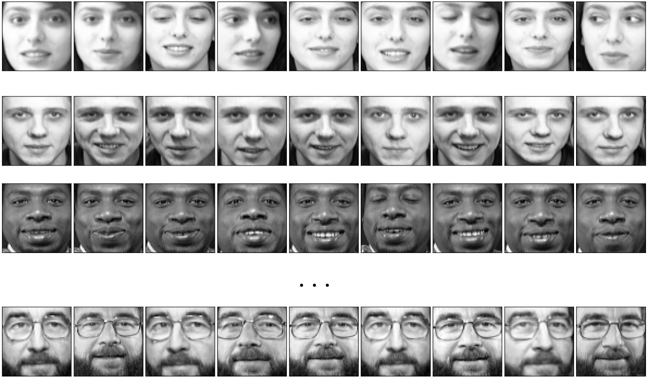
$\mathcal{S}_1 := \mathsf{span}\Bigg($  $\Bigg)$

Denoising via projection onto $\mathcal{S}_1$



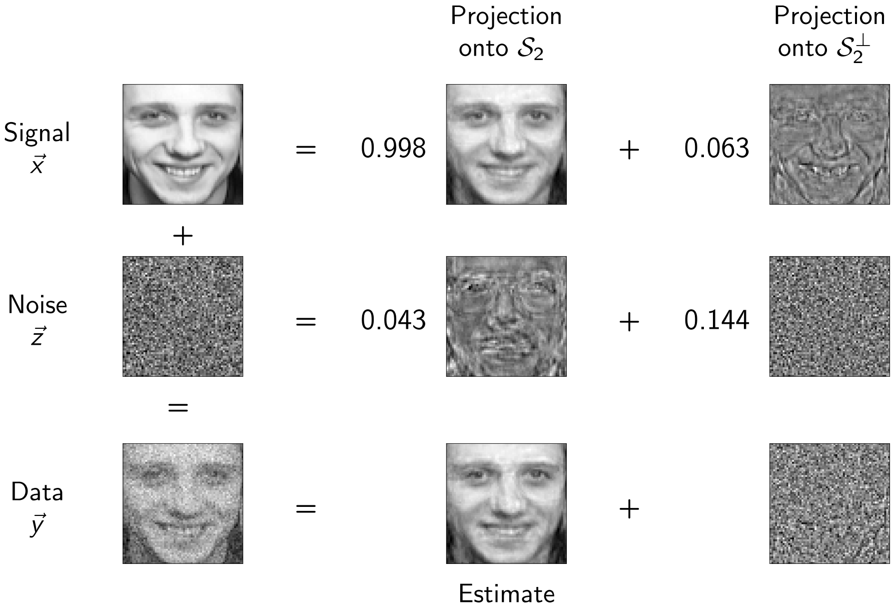
$\mathcal{S}_2$

$\mathcal{S}_2 := \text{span} \Bigg($



$\Bigg)$

Denoising via projection onto $\mathcal{S}_2$



$\mathcal{P}_{\mathcal{S}_1}\,\vec{z}$ and $\mathcal{P}_{\mathcal{S}_2}\,\vec{z}$

$\mathcal{P}_{\mathcal{S}_1}\,\vec{z}$

$\mathcal{P}_{\mathcal{S}_2}\,\vec{z}$

$$0.007 = \frac{||\mathcal{P}_{\mathcal{S}_1}\,\vec{z}||_2}{||\vec{x}||_2} < \frac{||\mathcal{P}_{\mathcal{S}_2}\,\vec{z}||_2}{||\vec{x}||_2} = 0.043$$

$$\frac{0.043}{0.007} = 6.14 \approx \sqrt{\frac{\dim(\mathcal{S}_2)}{\dim(\mathcal{S}_1)}} \qquad \text{(not a coincidence)}$$

$\mathcal{P}_{\mathcal{S}_1^\perp}\,\vec{x}$ and $\mathcal{P}_{\mathcal{S}_2^\perp}\,\vec{x}$

$\mathcal{P}_{\mathcal{S}_1^\perp}\,\vec{x}$

$\mathcal{P}_{\mathcal{S}_2^\perp}\,\vec{x}$

$$0.063 = \frac{\left|\left|\mathcal{P}_{\mathcal{S}_2^\perp}\,\vec{x}\right|\right|_2}{||\vec{x}||_2} < \frac{\left|\left|\mathcal{P}_{\mathcal{S}_1^\perp}\,\vec{x}\right|\right|_2}{||\vec{x}||_2} = 0.190$$

$\mathcal{P}_{\mathcal{S}_1}\,\vec{y}$ and $\mathcal{P}_{\mathcal{S}_2}\,\vec{y}$