Principal component analysis

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

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

Carlos Fernandez-Granda

Discussion

Covariance matrix

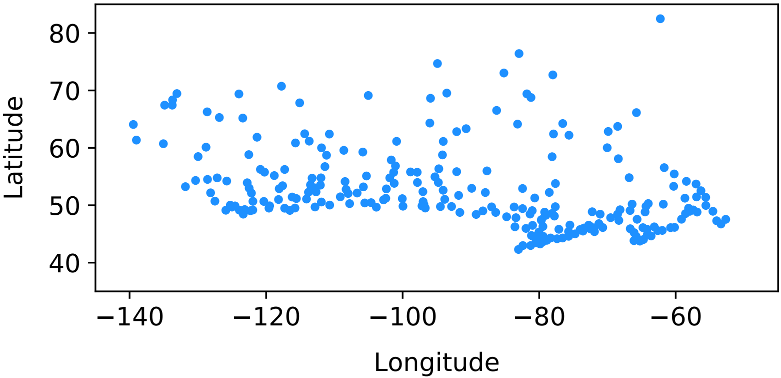
The spectral theorem

Principal component analysis

Dimensionality reduction via PCA

Gaussian random vectors

Motivation: Multidimensional data



Center of dataset

Probabilistic perspective: Data sampled from random vector $\tilde{x}$

What is the center of the dataset?

Center of dataset

Probabilistic perspective: Data sampled from random vector $\tilde{x}$

What is the center of the dataset?

Possible definition: Minimum difference to all the points on average

$$\text{Center} := \arg\min_{w \in \mathbb{R}^d} \mathrm{E}\left(||\tilde{x} - w||_2^2 \right)$$

Center of dataset

Probabilistic perspective: Data sampled from random vector $\tilde{x}$

What is the center of the dataset?

Possible definition: Minimum difference to all the points on average

$$
\begin{aligned}
\text{Center} &:= \arg \min_{w \in \mathbb{R}^d} \mathrm{E}\left(||\tilde{x} - w||_2^2 \right) \\
&= \arg \min_{w \in \mathbb{R}^d} \sum_{j=1}^{d} \mathrm{E}\left((\tilde{x}[j] - w[j])^2 \right)
\end{aligned}
$$

Center of dataset

Probabilistic perspective: Data sampled from random vector $\tilde{x}$

What is the center of the dataset?

Possible definition: Minimum difference to all the points on average

$$\begin{aligned}
\text{Center} &:= \arg\min_{w \in \mathbb{R}^d} \mathrm{E}\left(||\tilde{x} - w||_2^2\right) \\
&= \arg\min_{w \in \mathbb{R}^d} \sum_{j=1}^{d} \mathrm{E}\left((\tilde{x}[j] - w[j])^2\right) \\
&= \begin{bmatrix} \mathrm{E}(\tilde{x}[1]) \\ \cdots \\ \mathrm{E}(\tilde{x}[d]) \end{bmatrix}
\end{aligned}$$

Center of dataset

Probabilistic perspective: Data sampled from random vector $\tilde{x}$

What is the center of the dataset?

Possible definition: Minimum difference to all the points on average

$$
\begin{aligned}
\text{Center} &:= \arg\min_{w \in \mathbb{R}^d} \mathrm{E}\left(||\tilde{x} - w||_2^2 \right) \\
&= \arg\min_{w \in \mathbb{R}^d} \sum_{j=1}^{d} \mathrm{E}\left((\tilde{x}[j] - w[j])^2 \right) \\
&= \begin{bmatrix} \mathrm{E}(\tilde{x}[1]) \\ \cdots \\ \mathrm{E}(\tilde{x}[d]) \end{bmatrix} \\
&= \mathrm{E}(\tilde{x})
\end{aligned}
$$

Center of dataset

In practice, we have a dataset of n d-dimensional vectors
$\mathcal{X} := \{x_1, \ldots, x_n\}$

What is the center of the dataset?

Center of dataset

In practice, we have a dataset of n d-dimensional vectors $\mathcal{X} := \{x_1, \ldots, x_n\}$

What is the center of the dataset?

Reasonable choise: Sample mean

$$\mathrm{av}(\mathcal{X}) := \frac{1}{n}\sum_{i=1}^{n} x_i$$

Geometric interpretation

$$\text{Geometric center} := \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} ||x_i - w||_2^2$$

Geometric interpretation

$$
\begin{aligned}
\text{Geometric center} &:= \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} ||x_i - w||_2^2 \\
&= \arg\min_{w \in \mathbb{R}^d} \sum_{j=1}^{d} \sum_{i=1}^{n} (x_i[j] - w[j])^2
\end{aligned}
$$

Geometric interpretation

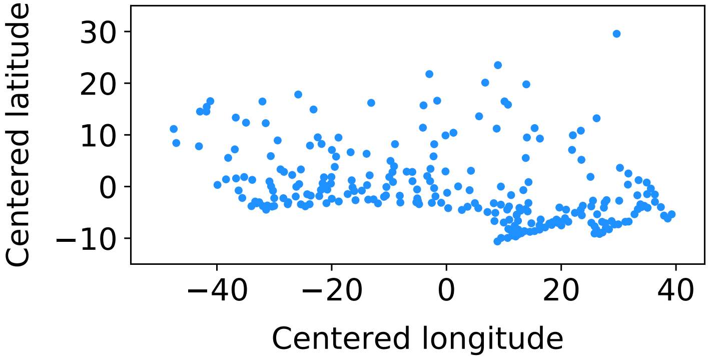
$$
\begin{aligned}
\text{Geometric center} &:= \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} ||x_i - w||_2^2 \\
&= \arg\min_{w \in \mathbb{R}^d} \sum_{j=1}^{d} \sum_{i=1}^{n} (x_i[j] - w[j])^2 \\
&= \begin{bmatrix} \frac{1}{n} \sum_i x_i[1] \\ \cdots \\ \frac{1}{n} \sum_i x_i[1] \end{bmatrix}
\end{aligned}
$$

Geometric interpretation

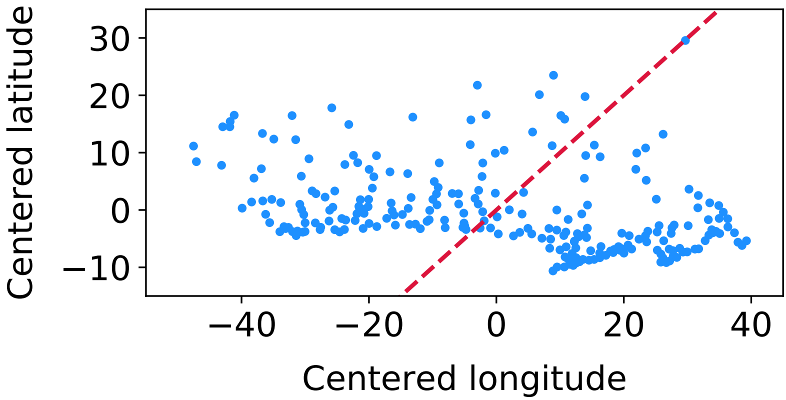
$$
\begin{aligned}
\text{Geometric center} &:= \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} ||x_i - w||_2^2 \\
&= \arg\min_{w \in \mathbb{R}^d} \sum_{j=1}^{d} \sum_{i=1}^{n} (x_i[j] - w[j])^2 \\
&= \begin{bmatrix} \frac{1}{n} \sum_i x_i[1] \\ \cdots \\ \frac{1}{n} \sum_i x_i[1] \end{bmatrix} \\
&= \text{av}(\mathcal{X})
\end{aligned}
$$

Centering

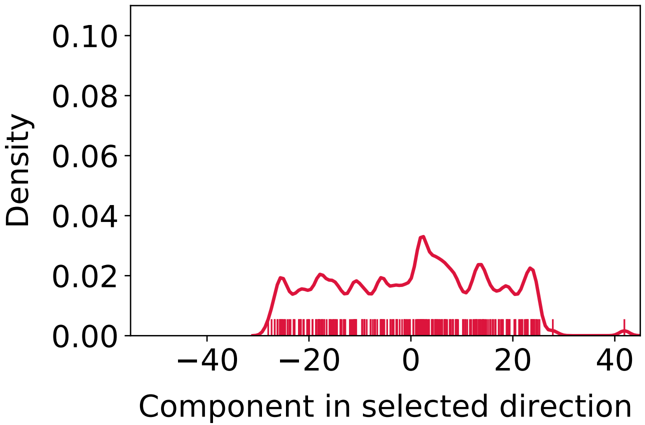
$$c(x_i) := x_i - \mathrm{av}(\mathcal{X})$$



Projection onto a fixed direction



Projection onto a fixed direction



Variance in direction of a fixed vector v

Variance in direction of a fixed vector v

$$\mathrm{Var}\left(v^T \tilde{x}\right)$$

Variance in direction of a fixed vector v

$$\mathrm{Var}\left(v^T\tilde{x}\right) = \mathrm{E}\left((v^T\tilde{x} - \mathrm{E}(v^T\tilde{x}))^2\right)$$

Variance in direction of a fixed vector v

$$\begin{aligned}\mathrm{Var}\left(v^T\tilde{x}\right) &= \mathrm{E}\left((v^T\tilde{x} - \mathrm{E}(v^T\tilde{x}))^2\right) \\ &= \mathrm{E}\left((v^T c(\tilde{x}))^2\right)\end{aligned}$$

Variance in direction of a fixed vector v

$$\begin{aligned}
\mathrm{Var}\left(v^T\tilde{x}\right) &= \mathrm{E}\left(\left(v^T\tilde{x} - \mathrm{E}(v^T\tilde{x})\right)^2\right) \\
&= \mathrm{E}\left(\left(v^T c(\tilde{x})\right)^2\right) \\
&= v^T \mathrm{E}\left(c(\tilde{x})c(\tilde{x})^T\right) v
\end{aligned}$$

Covariance matrix

The covariance matrix of a random vector $\tilde{x}$ is defined as

$$\Sigma_{\tilde{x}} := \mathrm{E}\left(c(\tilde{x})c(\tilde{x})^T\right)$$

$$= \begin{bmatrix} \mathrm{Var}\left(\tilde{x}[1]\right) & \mathrm{Cov}\left(\tilde{x}[1], \tilde{x}[2]\right) & \cdots & \mathrm{Cov}\left(\tilde{x}[1], \tilde{x}[d]\right) \\ \mathrm{Cov}\left(\tilde{x}[1], \tilde{x}[2]\right) & \mathrm{Var}\left(\tilde{x}[2]\right) & \cdots & \mathrm{Cov}\left(\tilde{x}[2], \tilde{x}[d]\right) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}\left(\tilde{x}[1], \tilde{x}[d]\right) & \mathrm{Cov}\left(\tilde{x}[2], \tilde{x}[d]\right) & \cdots & \mathrm{Var}\left(\tilde{x}[d]\right) \end{bmatrix}$$

Variance in direction of a fixed vector v

$$
\begin{aligned}
\mathrm{Var}\left(v^T\tilde{x}\right) &= \mathrm{E}\left(\left(v^T\tilde{x} - \mathrm{E}(v^T\tilde{x})\right)^2\right) \\
&= \mathrm{E}\left(\left(v^T c(\tilde{x})\right)^2\right) \\
&= v^T \mathrm{E}\left(c(\tilde{x})c(\tilde{x})^T\right) v \\
&= v^T \Sigma_{\tilde{x}} v
\end{aligned}
$$

Sample covariance matrix

For a dataset $\mathcal{X} = \{x_1, \ldots, x_n\}$

$$\begin{aligned}
\Sigma_{\mathcal{X}} &:= \frac{1}{n} \sum_{i=1}^{n} c(x_i) c(x_i)^T \\
&= \begin{bmatrix}
\operatorname{var}(\mathcal{X}[1]) & \operatorname{cov}(\mathcal{X}[1], \mathcal{X}[2]) & \cdots & \operatorname{cov}(\mathcal{X}[1], \mathcal{X}[d]) \\
\operatorname{cov}(\mathcal{X}[1], \mathcal{X}[2]) & \operatorname{var}(\mathcal{X}[2]) & \cdots & \operatorname{cov}(\mathcal{X}[2], \mathcal{X}[d]) \\
\vdots & \vdots & \ddots & \vdots \\
\operatorname{cov}(\mathcal{X}[1], \mathcal{X}[d]) & \operatorname{cov}(\mathcal{X}[2], \mathcal{X}[d]) & \cdots & \operatorname{var}(\mathcal{X}[d])
\end{bmatrix}
\end{aligned}$$

where $\mathcal{X}_i := \{x[i]_1, \ldots, x[i]_n\}$

Sample variance in direction of a fixed vector $\boldsymbol{v}$

$$\operatorname{var}\left(\mathcal{P}_{v}\,\mathcal{X}\right)$$

Sample variance in direction of a fixed vector v

$$\operatorname{var}\left(\mathcal{P}_v\,\mathcal{X}\right) := \frac{1}{n}\sum_{i=1}^{n}\left(v^T x_i - \operatorname{av}\left(\mathcal{P}_v\,\mathcal{X}\right)\right)^2$$

Sample variance in direction of a fixed vector v

$$\begin{aligned}
\operatorname{var}\left(\mathcal{P}_v\,\mathcal{X}\right) &:= \frac{1}{n}\sum_{i=1}^{n}\left(v^T x_i - \operatorname{av}\left(\mathcal{P}_v\,\mathcal{X}\right)\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(v^T\left(x_i - \operatorname{av}\left(\mathcal{X}\right)\right)\right)^2
\end{aligned}$$

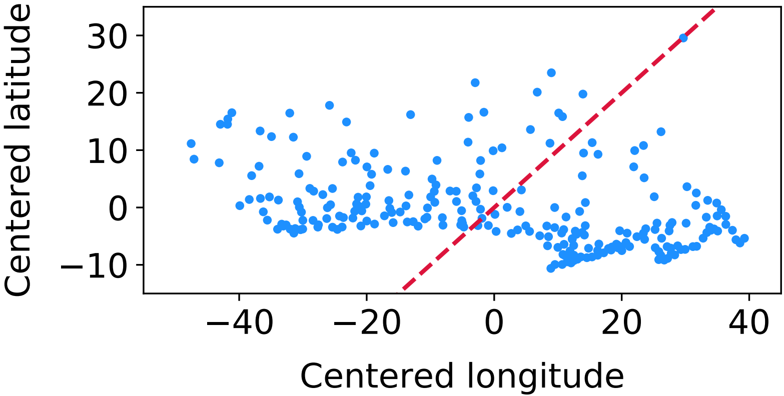
Sample variance in direction of a fixed vector v

$$
\begin{aligned}
\operatorname{var}\left(\mathcal{P}_v\,\mathcal{X}\right) &:= \frac{1}{n}\sum_{i=1}^{n}\left(v^T x_i - \operatorname{av}\left(\mathcal{P}_v\,\mathcal{X}\right)\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(v^T\left(x_i - \operatorname{av}\left(\mathcal{X}\right)\right)\right)^2 \\
&= v^T\left(\frac{1}{n}\sum_{i=1}^{n} c(x_i)c(x_i)^T\right)v
\end{aligned}
$$

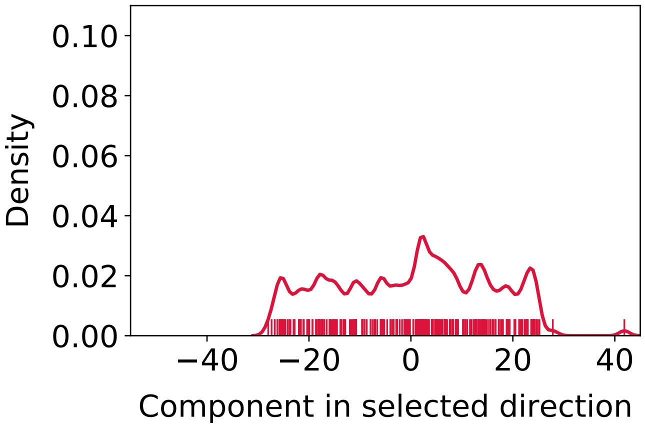
Sample variance in direction of a fixed vector v

$$\begin{aligned}
\operatorname{var}(\mathcal{P}_v \mathcal{X}) &:= \frac{1}{n}\sum_{i=1}^{n}\left(v^T x_i - \operatorname{av}(\mathcal{P}_v \mathcal{X})\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(v^T (x_i - \operatorname{av}(\mathcal{X}))\right)^2 \\
&= v^T \left(\frac{1}{n}\sum_{i=1}^{n} c(x_i)c(x_i)^T\right) v \\
&= v^T \Sigma_{\mathcal{X}} v
\end{aligned}$$

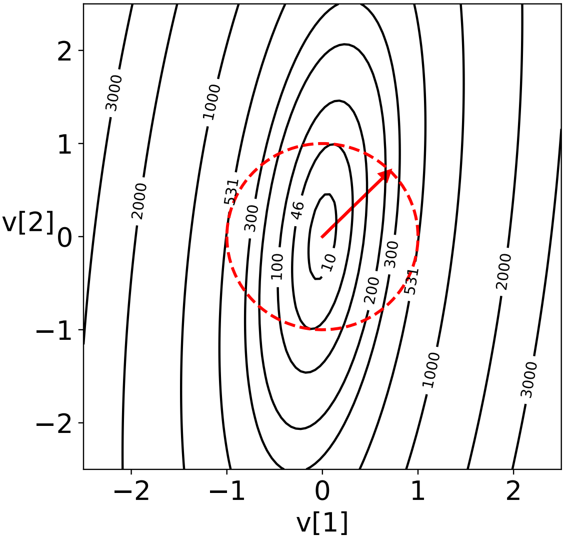
Sample variance = 229 (sample std = 15.1)



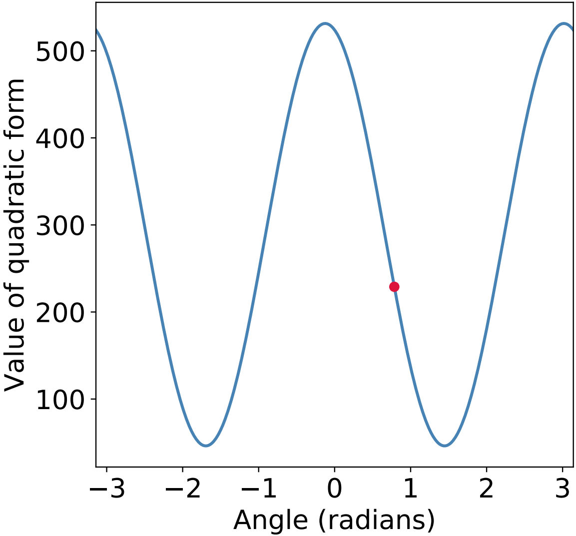
Sample variance = 229 (sample std = 15.1)



$f(v) := v^T \Sigma_{\mathcal{X}} v$ for $||v||_2 = 1$



$f(v) := v^T \Sigma_{\mathcal{X}} v$ for $||v||_2 = 1$



Quadratic form

Function $f : \mathbb{R}^d \to \mathbb{R}$ defined by

$$f(x) := x^T A x$$

where A is a $d \times d$ symmetric matrix

Generalization of quadratic functions to multiple dimensions

Quadratic form

Function $f : \mathbb{R}^d \to \mathbb{R}$ defined by

$$f(x) := x^T A x$$

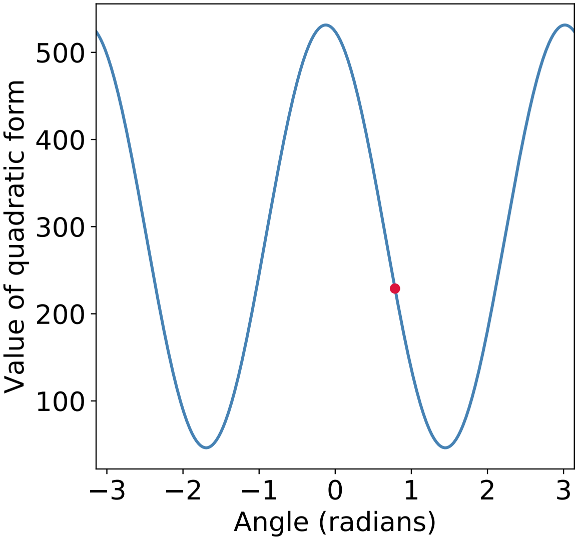
where A is a $d \times d$ symmetric matrix

Generalization of quadratic functions to multiple dimensions

Goal: Study quadratic forms when $||v||_2 = 1$

Motivation: If A is a covariance matrix, f encodes directional variance

Does the function necessarily reach a maximum?



Does the function necessarily reach a maximum? Yes

- The function is continuous (second-order polynomial)

Does the function necessarily reach a maximum? Yes

- The function is continuous (second-order polynomial)
- Unit sphere is closed and bounded (contains all limit points)

Does the function necessarily reach a maximum? Yes

- The function is continuous (second-order polynomial)
- Unit sphere is closed and bounded (contains all limit points)
- Image of unit sphere is also closed and bounded

Does the function necessarily reach a maximum? Yes

- The function is continuous (second-order polynomial)
- Unit sphere is closed and bounded (contains all limit points)
- Image of unit sphere is also closed and bounded
- Image cannot grow towards limit it does not contain

Does the function necessarily reach a maximum? Yes

For any symmetric matrix $A \in \mathbb{R}^{d \times d}$, there exists $u_1 \in \mathbb{R}^d$ such that

$$u_1 = \arg\max_{||x||_2 = 1} x^T A x$$

Directional derivative

For any differentiable $f : \mathbb{R}^d \to \mathbb{R}$ and any $v \in \mathbb{R}^d$ such that $||v||_2 = 1$

$$f'_v(x) := \lim_{h \to 0} \frac{f(x + hv) - f(x)}{h}$$
$$= \langle \nabla f(x), u \rangle$$

If $f'_v(x) > 0$, then $f(x + \epsilon v) > f(x)$ for sufficiently small $\epsilon > 0$

Characterizing maximum of quadratic form

At the maximum u_1, we cannot have

$$\begin{aligned} f'_v(u_1) &= \langle \nabla f(u_1), v \rangle \\ &\neq 0 \end{aligned}$$

for any v such that $u_1 + \epsilon v$ is in the constraint set

Characterizing maximum of quadratic form

At the maximum u_1, we cannot have

$$\begin{aligned} f'_v(u_1) &= \langle \nabla f(u_1), v \rangle \\ &\neq 0 \end{aligned}$$

for any v such that $u_1 + \epsilon v$ is in the constraint set

Wait a minute, *can* $u_1 + \epsilon v$ *be in our constraint set?*

Tangent hyperplane

Unit sphere is level surface of

$$g(x) := x^T x$$

$x + v$ is in the tangent plane of g at x if

$$\nabla g(x)^T v = 0$$

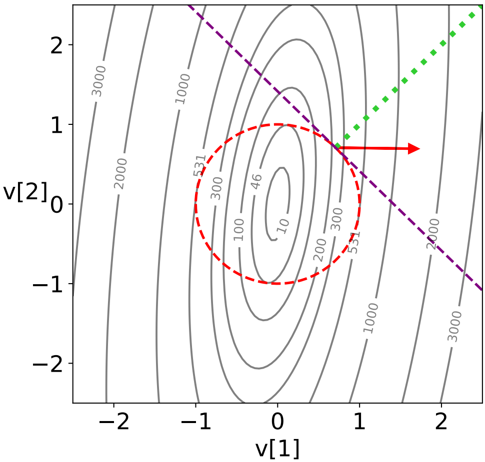
If v is in the tangent plane, then $g'_v(x) = 0$, so

$$g(x + \epsilon v) \approx g(x),$$

i.e. $x + \epsilon v$ is arbitrarily close to the level surface

Can this point be a maximum of the quadratic form?

Red arrow = gradient of quadratic form
Green line = gradient of $g(x) := x^T x$



Characterizing maximum of quadratic form

If

$$\langle \nabla f(u_1), v \rangle \neq 0$$

for some v in the tangent plane, then

$$f(u_1 + \epsilon v) > f(u_1)$$

for a point that is *almost* on the unit sphere

Characterizing maximum of quadratic form

If

$$\langle \nabla f(u_1), v \rangle \neq 0$$

for some v in the tangent plane, then

$$f(u_1 + \epsilon v) > f(u_1)$$

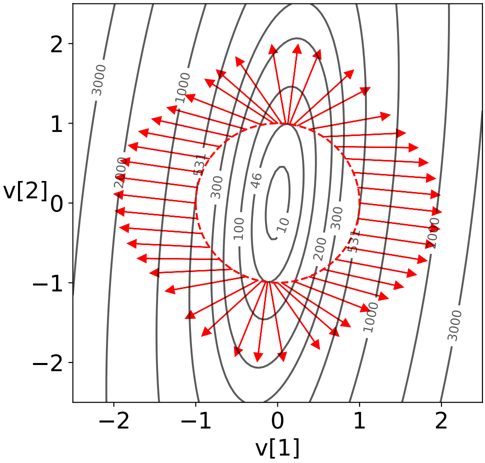
for a point that is *almost* on the unit sphere

Since f is continuous there exists a y on the sphere such that

$$f(y) \approx f(u_1 + \epsilon v) > f(u_1)$$

Where is the maximum?

Red arrow = gradient of quadratic form



Characterizing maximum of quadratic form

We need

$$\langle \nabla f(u_1), v \rangle = 0$$

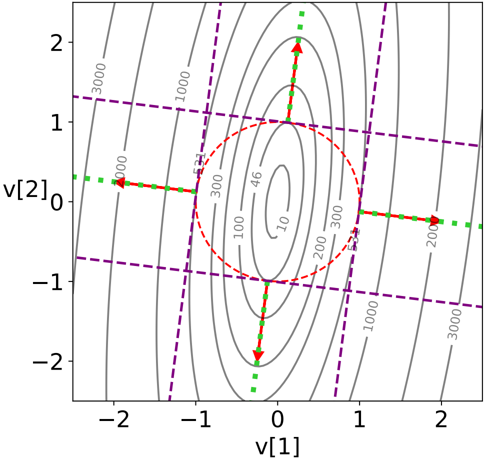
for *all* v in the tangent plane

Equivalent to $\nabla f(u_1) = \lambda_1 \nabla g(u_1)$ for some $\lambda_1 \in \mathbb{R}$. Then

$$\begin{aligned} \langle \nabla f(u_1), v \rangle &= \lambda_1 \langle \nabla g(u_1), v \rangle \\ &= 0 \end{aligned}$$

Maxima and minima satisfy $\nabla f(u_1) = \lambda_1 \nabla g(u_1)$

Red arrow = gradient of quadratic form
Green line = gradient of $g(x) := x^T x$



Conclusion

Maximum satisfies $\nabla f(u_1) = \lambda_1 \nabla g(u_1)$

$$\begin{aligned}\nabla f(x) &= \nabla x^T A x \\ &= \end{aligned}$$

$$\begin{aligned}\nabla g(x) &= \nabla x^T x \\ &= \end{aligned}$$

Conclusion

Maximum satisfies $\nabla f\left(u_1\right) = \lambda_1 \nabla g\left(u_1\right)$

$$\begin{aligned}
\nabla f\left(x\right) &= \nabla x^T A x \\
&= 2Ax
\end{aligned}$$

$$\begin{aligned}
\nabla g\left(x\right) &= \nabla x^T x \\
&=
\end{aligned}$$

Conclusion

Maximum satisfies $\nabla f\left(u_1\right) = \lambda_1 \nabla g\left(u_1\right)$

$$\begin{aligned}\nabla f\left(x\right) &= \nabla x^T A x \\ &= 2Ax\end{aligned}$$

$$\begin{aligned}\nabla g\left(x\right) &= \nabla x^T x \\ &= 2x\end{aligned}$$

Conclusion

Maximum satisfies $\nabla f(u_1) = \lambda_1 \nabla g(u_1)$

$$\begin{aligned}\nabla f(x) &= \nabla x^T A x \\ &= 2Ax\end{aligned}$$

$$\begin{aligned}\nabla g(x) &= \nabla x^T x \\ &= 2x\end{aligned}$$

so $Au_1 = \lambda_1 u_1$, i.e. u_1 is an eigenvector!

Conclusion

For any symmetric $A \in \mathbb{R}^{d \times d}$,

$$u_1 := \arg \max_{||x||_2 = 1} x^T A x$$

is an eigenvector of A. There exists $\lambda_1 \in \mathbb{R}$ such that

$$A u_1 = \lambda_1 u_1$$

Value of the maximum

We have

$$\max_{||x||_2=1} x^T A x = u_1^T A u_1$$

$$=$$

Value of the maximum

We have

$$\begin{aligned}\max_{||x||_2=1} x^T A x &= u_1^T A u_1 \\ &= \lambda_1\end{aligned}$$

Are there more eigenvectors?

Think about $A \in \mathbb{R}^{3\times 3}$

We know u_1 attains maximum

Are there more eigenvectors?

Think about $A \in \mathbb{R}^{3\times 3}$

We know u_1 attains maximum

What happens on plane orthogonal to u_1?

Without loss of generality assume $u_1 = e_3$

Constraint set?

Quadratic function?

Are there more eigenvectors?

Think about $A \in \mathbb{R}^{3\times 3}$

We know u_1 attains maximum

What happens on plane orthogonal to u_1?

Without loss of generality assume $u_1 = e_3$

Constraint set? Circle

Quadratic function?

Are there more eigenvectors?

Think about $A \in \mathbb{R}^{3\times 3}$

We know u_1 attains maximum

What happens on plane orthogonal to u_1?

Without loss of generality assume $u_1 = e_3$

Constraint set? Circle

Quadratic function?

$$x^T A x = \begin{bmatrix} x[1] \\ x[2] \end{bmatrix}^T \begin{bmatrix} A[1,1] & A[1,2] \\ A[2,1] & A[2,2] \end{bmatrix} \begin{bmatrix} x[1] \\ x[2] \end{bmatrix}$$

Are there more eigenvectors?

Think about $A \in \mathbb{R}^{3\times 3}$

We know u_1 attains maximum

What happens on plane orthogonal to u_1?

Without loss of generality assume $u_1 = e_3$

Constraint set? Circle

Quadratic function?

$$x^T A x = \begin{bmatrix} x[1] \\ x[2] \end{bmatrix}^T \begin{bmatrix} A[1,1] & A[1,2] \\ A[2,1] & A[2,2] \end{bmatrix} \begin{bmatrix} x[1] \\ x[2] \end{bmatrix}$$

So there exists eigenvector u_2...

Spectral theorem

If $A \in \mathbb{R}^{d \times d}$ is symmetric, then it has an eigendecomposition

$$A = \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_d \end{bmatrix} \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix}^T,$$

Eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ are real

Eigenvectors $u_1, u_2, \ldots, u_n$ are real and orthogonal

Spectral theorem

$$\lambda_1 = \max_{||x||_2=1} x^T A x$$

$$u_1 = \arg\max_{||x||_2=1} x^T A x$$

$$\lambda_k = \max_{||x||_2=1, x \perp u_1, \ldots, u_{k-1}} x^T A x, \quad 2 \le k \le d$$

$$u_k = \arg\max_{||x||_2=1, x \perp u_1, \ldots, u_{k-1}} x^T A x, \quad 2 \le k \le d$$

How do we prove this?

How do we prove this?

Formalize intuition from 3×3 case through induction

Mathematical induction

If a statement $\mathcal{S}_d$ dependent on d satisfies:

- $\mathcal{S}_1$ holds (basis)
- If $\mathcal{S}_{d-1}$ holds then $\mathcal{S}_d$ holds (step)

Then $\mathcal{S}_d$ is true for all natural numbers $d = 1, 2, \ldots$

Basis

For $d = 1$ what is u_1 and λ_1?

Step

We know u_1 exists and satisfies $Au_1 = \lambda_1 u_1$

Let us consider action of A on orthogonal complement of u_1

Step

We know u_1 exists and satisfies $Au_1 = \lambda_1 u_1$

Let us consider action of A on orthogonal complement of u_1

We want matrix A' such that

$$A' u_1 = 0$$

$$A' x = x \quad \text{if } x \perp u_1$$

Step

We know u_1 exists and satisfies $Au_1 = \lambda_1 u_1$

Let us consider action of A on orthogonal complement of u_1

We want matrix A' such that

$$A' u_1 = 0$$

$$A' x = x \quad \text{if } x \perp u_1$$

$A - \lambda_1 u_1 u_1^T$ works

Step

We want to apply assumption about $d-1 \times d-1$ matrices

We need to "compress" $A - \lambda_1 u_1 u_1^T$

Step

We want to apply assumption about $d-1 \times d-1$ matrices

We need to "compress" $A - \lambda_1 u_1 u_1^T$

Let $V_\perp \in \mathbb{R}^{d-1 \times d}$ contain orthonormal basis of $\text{span}(u_1)^\perp$

$V_\perp V_\perp^T$ is projection matrix

Step

We want to apply assumption about $d-1 \times d-1$ matrices

We need to "compress" $A - \lambda_1 u_1 u_1^T$

Let $V_\perp \in \mathbb{R}^{d-1 \times d}$ contain orthonormal basis of $\text{span}(u_1)^\perp$

$V_\perp V_\perp^T$ is projection matrix

$$V_\perp V_\perp^T (Au_1 - \lambda_1 u_1) V_\perp V_\perp^T =$$

Step

We want to apply assumption about $d - 1 \times d - 1$ matrices

We need to "compress" $A - \lambda_1 u_1 u_1^T$

Let $V_\perp \in \mathbb{R}^{d-1 \times d}$ contain orthonormal basis of span$(u_1)^\perp$

$V_\perp V_\perp^T$ is projection matrix

$$V_\perp V_\perp^T (Au_1 - \lambda_1 u_1) V_\perp V_\perp^T = A - \lambda_1 u_1 u_1^T$$

Step

We want to apply assumption about $d-1 \times d-1$ matrices

We need to "compress" $A - \lambda_1 u_1 u_1^T$

Let $V_\perp \in \mathbb{R}^{d-1 \times d}$ contain orthonormal basis of $\text{span}(u_1)^\perp$

$V_\perp V_\perp^T$ is projection matrix

$$V_\perp V_\perp^T (Au_1 - \lambda_1 u_1) V_\perp V_\perp^T = A - \lambda_1 u_1 u_1^T$$

We define symmetric $B := V_\perp^T (Au_1 - \lambda_1 u_1) V_\perp \in \mathbb{R}^{d-1 \times d-1}$

Step

By induction assumption there exist $\gamma_1, \ldots, \gamma_{d-1}$ and $w_1, \ldots, w_{d-1}$ such that

$$\gamma_1 = \max_{||y||_2=1} y^T B y$$

$$w_1 = \arg \max_{||y||_2=1} y^T B y$$

$$\gamma_k = \max_{||y||_2=1, y \perp w_1, \ldots, w_{k-1}} y^T B y, \quad 2 \leq k \leq d-2$$

$$w_k = \arg \max_{||y||_2=1, y \perp w_1, \ldots, w_{k-1}} y^T B y, \quad 2 \leq k \leq d-2$$

Step

For any $x \in \text{span}(u_1)^\perp$, $x = V_\perp y$ for some $y \in \mathbb{R}^{d-1}$

$$\max_{||x||_2=1, x \perp u_1} x^T A x =$$

Step

For any $x \in \text{span}(u_1)^\perp$, $x = V_\perp y$ for some $y \in \mathbb{R}^{d-1}$

$$\max_{||x||_2=1, x \perp u_1} x^T A x = \max_{||x||_2=1, x \perp u_1} x^T (A - \lambda_1 u_1 u_1^T) x$$

Step

For any $x \in \text{span}(u_1)^\perp$, $x = V_\perp y$ for some $y \in \mathbb{R}^{d-1}$

$$\begin{aligned}
\max_{||x||_2=1, x \perp u_1} x^T A x &= \max_{||x||_2=1, x \perp u_1} x^T (A - \lambda_1 u_1 u_1^T) x \\
&= \max_{||x||_2=1, x \perp u_1} x^T V_\perp V_\perp^T (A u_1 - \lambda_1 u_1) V_\perp V_\perp^T x
\end{aligned}$$

Step

For any $x \in \text{span}(u_1)^\perp$, $x = V_\perp y$ for some $y \in \mathbb{R}^{d-1}$

$$\begin{aligned}
\max_{||x||_2=1, x \perp u_1} x^T A x &= \max_{||x||_2=1, x \perp u_1} x^T (A - \lambda_1 u_1 u_1^T) x \\
&= \max_{||x||_2=1, x \perp u_1} x^T V_\perp V_\perp^T (A u_1 - \lambda_1 u_1) V_\perp V_\perp^T x \\
&= \max_{||y||_2=1} y^T B y
\end{aligned}$$

Step

For any $x \in \text{span}(u_1)^{\perp}$, $x = V_{\perp} y$ for some $y \in \mathbb{R}^{d-1}$

$$\begin{aligned}
\max_{||x||_2=1, x \perp u_1} x^T A x &= \max_{||x||_2=1, x \perp u_1} x^T (A - \lambda_1 u_1 u_1^T) x \\
&= \max_{||x||_2=1, x \perp u_1} x^T V_{\perp} V_{\perp}^T (A u_1 - \lambda_1 u_1) V_{\perp} V_{\perp}^T x \\
&= \max_{||y||_2=1} y^T B y \\
&= \gamma_1
\end{aligned}$$

Step

For any $x \in \text{span}(u_1)^\perp$, $x = V_\perp y$ for some $y \in \mathbb{R}^{d-1}$

$$
\begin{aligned}
\max_{||x||_2=1, x\perp u_1} x^T A x &= \max_{||x||_2=1, x\perp u_1} x^T(A - \lambda_1 u_1 u_1^T)x \\
&= \max_{||x||_2=1, x\perp u_1} x^T V_\perp V_\perp^T (Au_1 - \lambda_1 u_1) V_\perp V_\perp^T x \\
&= \max_{||y||_2=1} y^T B y \\
&= \gamma_1
\end{aligned}
$$

Inspired by this: $u_k := V_\perp w_{k-1}$ for $k = 2, \ldots, d$

Step

For any $x \in \text{span}(u_1)^\perp$, $x = V_\perp y$ for some $y \in \mathbb{R}^{d-1}$

$$\begin{aligned}
\max_{||x||_2=1, x\perp u_1} x^T A x &= \max_{||x||_2=1, x\perp u_1} x^T (A - \lambda_1 u_1 u_1^T) x \\
&= \max_{||x||_2=1, x\perp u_1} x^T V_\perp V_\perp^T (A u_1 - \lambda_1 u_1) V_\perp V_\perp^T x \\
&= \max_{||y||_2=1} y^T B y \\
&= \gamma_1
\end{aligned}$$

Inspired by this: $u_k := V_\perp w_{k-1}$ for $k = 2, \ldots, d$

$u_1, \ldots, u_d$ are orthonormal basis

Step: eigenvectors

$$Au_k =$$

Step: eigenvectors

$$Au_k = V_\perp V_\perp^T (A - \lambda_1 u_1 u_1^T) V_\perp V_\perp^T V_\perp w_{k-1}$$

Step: eigenvectors

$$
\begin{aligned}
Au_k &= V_\perp V_\perp^T (A - \lambda_1 u_1 u_1^T) V_\perp V_\perp^T V_\perp w_{k-1} \\
&= V_\perp B w_k
\end{aligned}
$$

Step: eigenvectors

$$\begin{aligned} Au_k &= V_\perp V_\perp^T (A - \lambda_1 u_1 u_1^T) V_\perp V_\perp^T V_\perp w_{k-1} \\ &= V_\perp B w_k \\ &= \gamma_{k-1} V_\perp w_{k-1} \end{aligned}$$

Step: eigenvectors

$$
\begin{aligned}
Au_k &= V_\perp V_\perp^T (A - \lambda_1 u_1 u_1^T) V_\perp V_\perp^T V_\perp w_{k-1} \\
&= V_\perp B w_k \\
&= \gamma_{k-1} V_\perp w_{k-1} \\
&= \lambda_k u_k
\end{aligned}
$$

Step: eigenvectors

$$
\begin{aligned}
Au_k &= V_\perp V_\perp^T (A - \lambda_1 u_1 u_1^T) V_\perp V_\perp^T V_\perp w_{k-1} \\
&= V_\perp B w_k \\
&= \gamma_{k-1} V_\perp w_{k-1} \\
&= \lambda_k u_k
\end{aligned}
$$

u_k is an eigenvector of A with eigenvalue $\lambda_k := \gamma_{k-1}$

Step

Let $x \in \mathrm{span}(u_1)^\perp$ be orthogonal to $u_{k'}$, where $2 \leq k' \leq d$

There is $y \in \mathbb{R}^{d-1}$ such that $x = V_\perp y$ and

$$w_{k'-1}^T y =$$

Step

Let $x \in \mathrm{span}(u_1)^\perp$ be orthogonal to $u_{k'}$, where $2 \leq k' \leq d$

There is $y \in \mathbb{R}^{d-1}$ such that $x = V_\perp y$ and

$$w_{k'-1}^T y = w_{k'}^T V_\perp^T V_\perp y$$

Step

Let $x \in \text{span}(u_1)^\perp$ be orthogonal to $u_{k'}$, where $2 \leq k' \leq d$

There is $y \in \mathbb{R}^{d-1}$ such that $x = V_\perp y$ and

$$\begin{aligned} w_{k'-1}^T y &= w_{k'}^T V_\perp^T V_\perp y \\ &= u_{k'}^T x \end{aligned}$$

Step

Let $x \in \text{span}(u_1)^\perp$ be orthogonal to $u_{k'}$, where $2 \leq k' \leq d$

There is $y \in \mathbb{R}^{d-1}$ such that $x = V_\perp y$ and

$$\begin{aligned} w_{k'-1}^T y &= w_{k'}^T V_\perp^T V_\perp y \\ &= u_{k'}^T x \\ &= 0 \end{aligned}$$

Step: eigenvalues

Let $x \in \text{span}(u_1)^\perp$ be orthogonal to $u_{k'}$, where $2 \leq k' \leq d$

There is $y \in \mathbb{R}^{d-1}$ such that $x = V_\perp y$ and

$$w_{k'-1}^T y = 0$$

$$\max_{||x||_2=1, x \perp u_1, \ldots, u_{k-1}} x^T A x =$$

Step: eigenvalues

Let $x \in \operatorname{span}(u_1)^\perp$ be orthogonal to $u_{k'}$, where $2 \leq k' \leq d$

There is $y \in \mathbb{R}^{d-1}$ such that $x = V_\perp y$ and

$$w_{k'-1}^T y = 0$$

$$\max_{||x||_2=1, x \perp u_1, \dots, u_{k-1}} x^T A x = \max_{||x||_2=1, x \perp u_1, \dots, u_{k-1}} x^T V_\perp V_\perp^T (A u_1 - \lambda_1 u_1) V_\perp V_\perp^T x$$

Step: eigenvalues

Let $x \in \text{span}(u_1)^\perp$ be orthogonal to $u_{k'}$, where $2 \le k' \le d$

There is $y \in \mathbb{R}^{d-1}$ such that $x = V_\perp y$ and

$$w_{k'-1}^T y = 0$$

$$\begin{aligned}
\max_{||x||_2=1, x \perp u_1, \dots, u_{k-1}} x^T A x &= \max_{||x||_2=1, x \perp u_1, \dots, u_{k-1}} x^T V_\perp V_\perp^T (Au_1 - \lambda_1 u_1) V_\perp V_\perp^T x \\
&= \max_{||y||_2=1, y \perp w_1, \dots, w_{k-2}} y^T B y
\end{aligned}$$

Step: eigenvalues

Let $x \in \text{span}(u_1)^\perp$ be orthogonal to $u_{k'}$, where $2 \le k' \le d$

There is $y \in \mathbb{R}^{d-1}$ such that $x = V_\perp y$ and

$$w_{k'-1}^T y = 0$$

$$\begin{aligned}
\max_{||x||_2=1, x\perp u_1,\ldots,u_{k-1}} x^T A x &= \max_{||x||_2=1, x\perp u_1,\ldots,u_{k-1}} x^T V_\perp V_\perp^T (Au_1 - \lambda_1 u_1) V_\perp V_\perp^T x \\
&= \max_{||y||_2=1, y\perp w_1,\ldots,w_{k-2}} y^T B y \\
&= \gamma_{k-1}
\end{aligned}$$

Step: eigenvalues

Let $x \in \text{span}(u_1)^\perp$ be orthogonal to $u_{k'}$, where $2 \leq k' \leq d$

There is $y \in \mathbb{R}^{d-1}$ such that $x = V_\perp y$ and

$$w_{k'-1}^T y = 0$$

$$\begin{aligned}
\max_{||x||_2=1, x \perp u_1, \ldots, u_{k-1}} x^T A x &= \max_{||x||_2=1, x \perp u_1, \ldots, u_{k-1}} x^T V_\perp V_\perp^T (Au_1 - \lambda_1 u_1) V_\perp V_\perp^T x \\
&= \max_{||y||_2=1, y \perp w_1, \ldots, w_{k-2}} y^T B y \\
&= \gamma_{k-1} \\
&= \lambda_k
\end{aligned}$$

Spectral theorem

If $A \in \mathbb{R}^{d \times d}$ is symmetric, then it has an eigendecomposition

$$A = \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \cdots & \\ 0 & 0 & \cdots & \lambda_d \end{bmatrix} \begin{bmatrix} u_1 & u_2 & \cdots & u_d \end{bmatrix}^T,$$

Eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ are real

Eigenvectors u_1, u_2, ..., u_n are real and orthogonal

Variance in direction of a fixed vector v

If random vector $\tilde{x}$ has covariance matrix $\Sigma_{\tilde{x}}$

$$\mathrm{Var}\left(v^T \tilde{x}\right) = v^T \Sigma_{\tilde{x}} v$$

Principal directions

Let u_1, ..., u_d, and $\lambda_1 > \ldots > \lambda_d$ be the eigenvectors/eigenvalues of $\Sigma_{\tilde{x}}$

$$\lambda_1 = \max_{||v||_2=1} \mathrm{Var}(v^T \tilde{x})$$

$$u_1 = \arg\max_{||v||_2=1} \mathrm{Var}(v^T \tilde{x})$$

$$\lambda_k = \max_{||v||_2=1, v \perp u_1, \ldots, u_{k-1}} \mathrm{Var}(v^T \tilde{x}), \quad 2 \leq k \leq d$$

$$u_k = \arg\max_{||v||_2=1, v \perp u_1, \ldots, u_{k-1}} \mathrm{Var}(v^T \tilde{x}), \quad 2 \leq k \leq d$$

Principal components

Let $c(\tilde{x}) := \tilde{x} - \mathrm{E}(\tilde{x})$

$$\widetilde{pc}[i] := u_i^T c(\tilde{x}), \quad 1 \leq i \leq d$$

is the ith principal component

$$\mathrm{Var}(\widetilde{pc}[i]) :=$$

Principal components

Let $c(\tilde{x}) := \tilde{x} - \mathrm{E}(\tilde{x})$

$$\widetilde{pc}[i] := u_i^T c(\tilde{x}), \quad 1 \leq i \leq d$$

is the ith principal component

$$\mathrm{Var}(\widetilde{pc}[i]) := \lambda_i, \quad 1 \leq i \leq d$$

Principal components are uncorrelated

$$\mathrm{E}(\widetilde{pc}[i]\widetilde{pc}[j]) =$$

Principal components are uncorrelated

$$\mathrm{E}(\widetilde{pc}[i]\widetilde{pc}[j]) = \mathrm{E}(u_i^T c(\tilde{x}) u_j^T c(\tilde{x}))$$

Principal components are uncorrelated

$$\begin{aligned}
\mathrm{E}(\widetilde{pc}[i]\widetilde{pc}[j]) &= \mathrm{E}(u_i^T c(\tilde{x}) u_j^T c(\tilde{x})) \\
&= u_i^T \mathrm{E}(c(\tilde{x}) c(\tilde{x})^T) u_j
\end{aligned}$$

Principal components are uncorrelated

$$\begin{aligned}
\mathrm{E}(\widetilde{pc}[i]\widetilde{pc}[j]) &= \mathrm{E}(u_i^T c(\tilde{x}) u_j^T c(\tilde{x})) \\
&= u_i^T \mathrm{E}(c(\tilde{x}) c(\tilde{x})^T) u_j \\
&= u_i^T \Sigma_{\tilde{x}} u_j
\end{aligned}$$

Principal components are uncorrelated

$$
\begin{aligned}
\mathrm{E}(\widetilde{pc}[i]\widetilde{pc}[j]) &= \mathrm{E}(u_i^T c(\tilde{x}) u_j^T c(\tilde{x})) \\
&= u_i^T \mathrm{E}(c(\tilde{x}) c(\tilde{x})^T) u_j \\
&= u_i^T \Sigma_{\tilde{x}} u_j \\
&= \lambda_i u_i^T u_j
\end{aligned}
$$

Principal components are uncorrelated

$$
\begin{aligned}
\mathrm{E}(\widetilde{pc}[i]\widetilde{pc}[j]) &= \mathrm{E}(u_i^T c(\tilde{x}) u_j^T c(\tilde{x})) \\
&= u_i^T \mathrm{E}(c(\tilde{x}) c(\tilde{x})^T) u_j \\
&= u_i^T \Sigma_{\tilde{x}} u_j \\
&= \lambda_i u_i^T u_j \\
&= 0
\end{aligned}
$$

Principal components

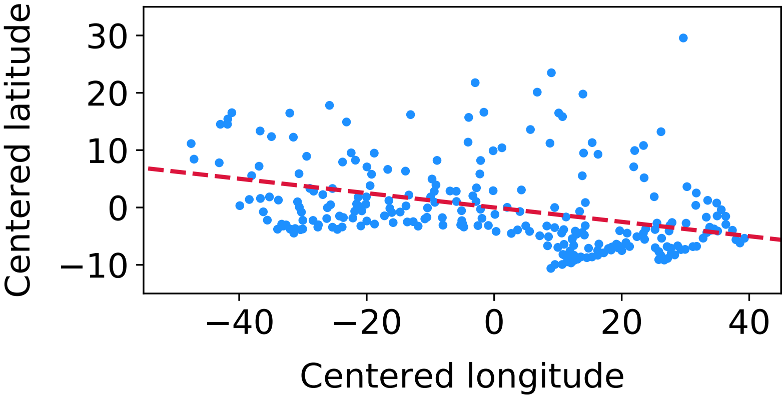
For dataset $\mathcal{X}$ containing $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$

1. Compute sample covariance matrix $\Sigma_{\mathcal{X}}$
2. Eigendecomposition of $\Sigma_{\mathcal{X}}$ yields principal directions u_1, ..., u_d
3. Center the data and compute principal components

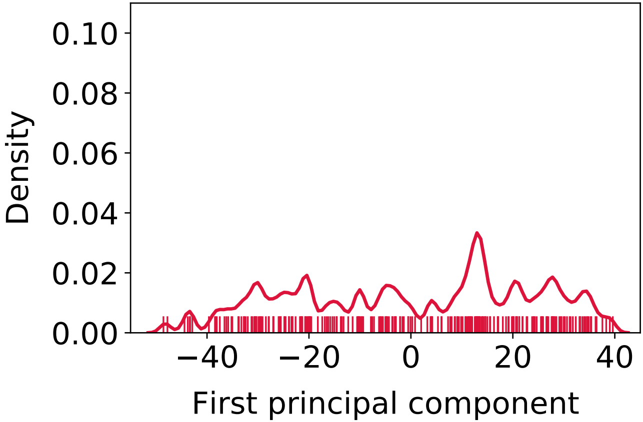
$$pc_i[j] := u_j^T c(x_i), \quad 1 \leq i \leq n, \; 1 \leq j \leq d,$$

where $c(x_i) := x_i - \text{av}(\mathcal{X})$

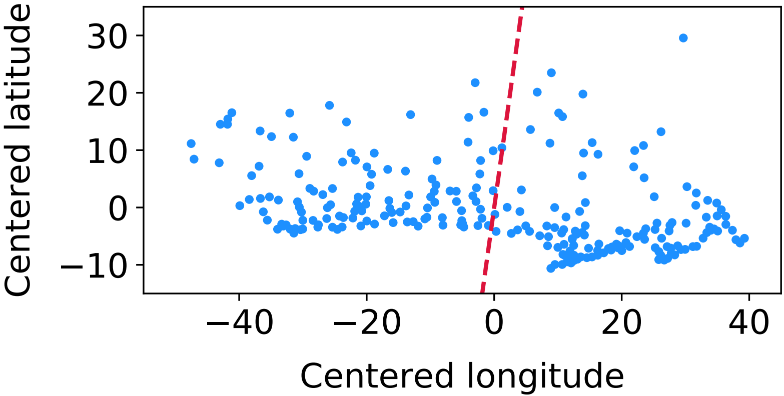
First principal direction



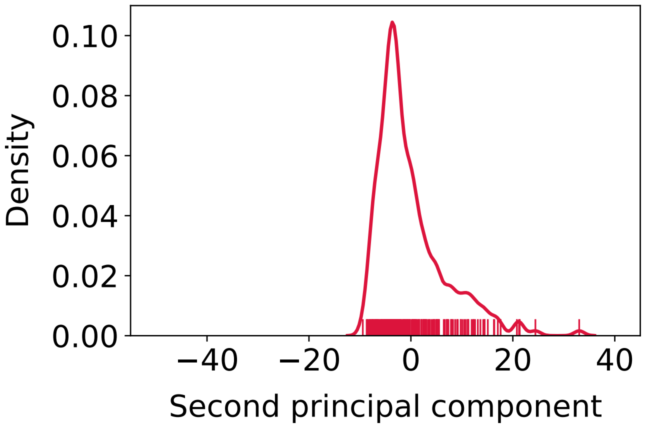
First principal component



Second principal direction



Second principal component



Sample variance in direction of a fixed vector v

$$\operatorname{var}\left(\mathcal{P}_v \mathcal{X}\right) = v^T \Sigma_{\mathcal{X}} v$$

Principal directions

Let $u_1, \ldots, u_d$, and $\lambda_1 > \ldots > \lambda_d$ be the eigenvectors/eigenvalues of $\Sigma_{\mathcal{X}}$

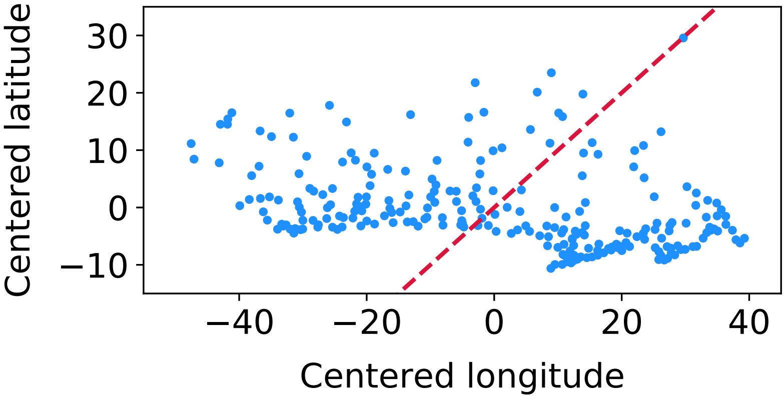
$$\lambda_1 = \max_{||v||_2=1} \operatorname{var}\left(\mathcal{P}_v\, \mathcal{X}\right)$$

$$u_1 = \arg\max_{||v||_2=1} \operatorname{var}\left(\mathcal{P}_v\, \mathcal{X}\right)$$

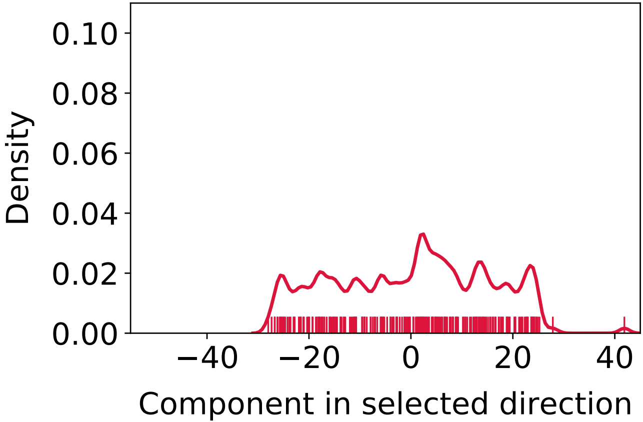
$$\lambda_k = \max_{||v||_2=1, v \perp u_1, \ldots, u_{k-1}} \operatorname{var}\left(\mathcal{P}_v\, \mathcal{X}\right), \quad 2 \leq k \leq d$$

$$u_k = \arg\max_{||v||_2=1, v \perp u_1, \ldots, u_{k-1}} \operatorname{var}\left(\mathcal{P}_v\, \mathcal{X}\right), \quad 2 \leq k \leq d$$

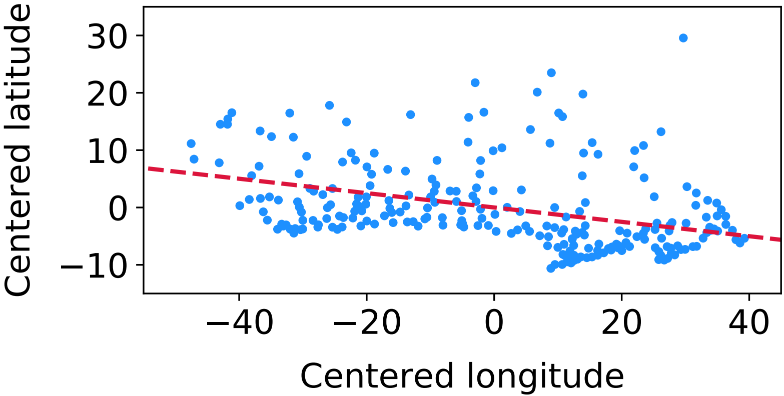
Sample variance = 229 (sample std = 15.1)



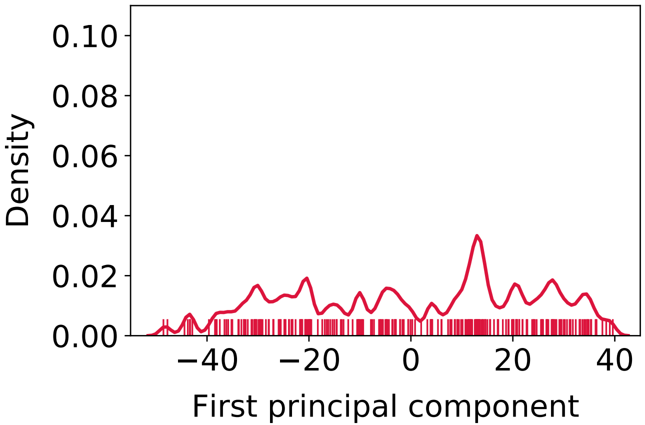
Sample variance = 229 (sample std = 15.1)



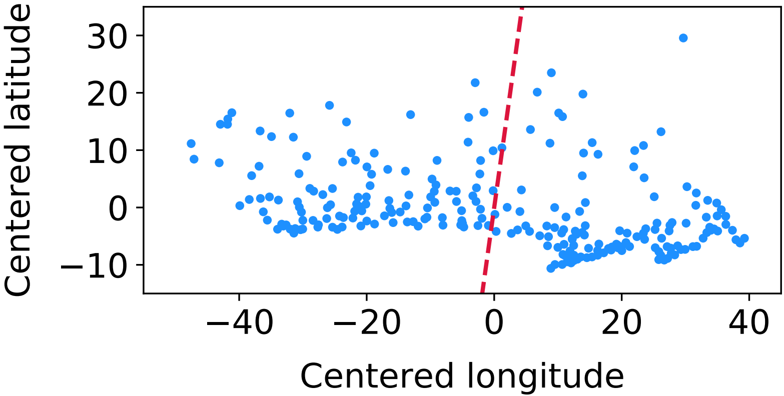
Sample variance = 531 (sample std = 23.1)



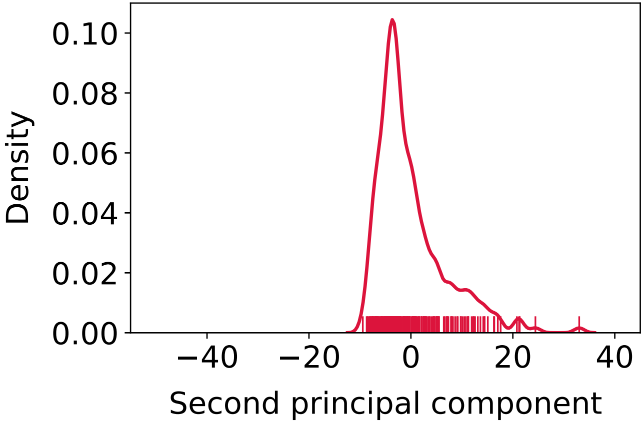
Sample variance = 531 (sample std = 23.1



Sample variance = 46.2 (sample std = 6.80)



Sample variance = 46.2 (sample std = 6.80)

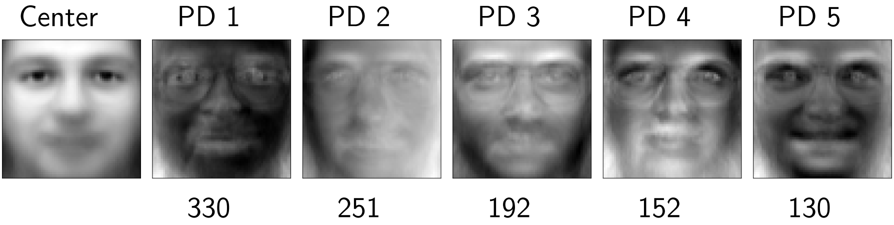


PCA of faces

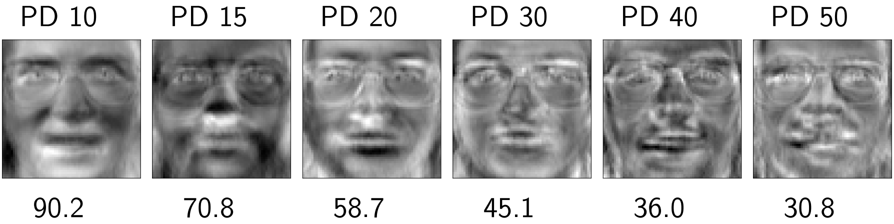
Data set of 400 64×64 images from 40 subjects (10 per subject)

Each face is vectorized and interpreted as a vector in $\mathbb{R}^{4096}$

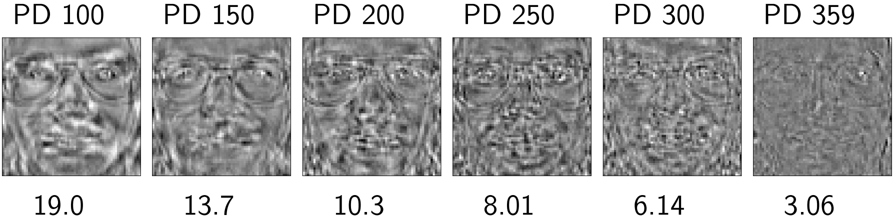
PCA of faces



PCA of faces



PCA of faces



Covariance matrix

The spectral theorem

Principal component analysis

Dimensionality reduction via PCA

Gaussian random vectors

Dimensionality reduction

Data with a large number of features can be difficult to analyze or process

Dimensionality reduction is a useful preprocessing step

If data are modeled as vectors in $\mathbb{R}^p$ we can reduce the dimension by projecting onto $\mathbb{R}^k$, where $k < p$

For orthogonal projections, the new representation is $\langle v_1, x \rangle$, $\langle v_2, x \rangle$, ..., $\langle v_k, x \rangle$ for a basis v_1, ..., v_k of the subspace that we project on

Problem: How do we choose the subspace?

Dimensionality reduction

Data with a large number of features can be difficult to analyze or process

Dimensionality reduction is a useful preprocessing step

If data are modeled as vectors in $\mathbb{R}^p$ we can reduce the dimension by projecting onto $\mathbb{R}^k$, where $k < p$

For orthogonal projections, the new representation is $\langle v_1, x \rangle$, $\langle v_2, x \rangle$, ..., $\langle v_k, x \rangle$ for a basis v_1, ..., v_k of the subspace that we project on

Problem: How do we choose the subspace?

Possible criterion: Capture as much sample variance as possible

Captured variance

For any orthonormal $v_1, \ldots, v_k$

$$\sum_{i=1}^{k} \operatorname{var}(\mathcal{P}_{v_i}\, \mathcal{X}) =$$

Captured variance

For any orthonormal $v_1, \ldots, v_k$

$$\sum_{i=1}^{k} \operatorname{var}(\mathcal{P}_{v_i}\, \mathcal{X}) = \sum_{i=1}^{k} \frac{1}{n} \sum_{j=1}^{n} {v_i}^T c(x_j) c(x_j)^T v_i$$

Captured variance

For any orthonormal $v_1, \ldots, v_k$

$$\begin{aligned}
\sum_{i=1}^{k} \text{var}(\mathcal{P}_{v_i}\, \mathcal{X}) &= \sum_{i=1}^{k} \frac{1}{n} \sum_{j=1}^{n} {v_i}^T c(x_j) c(x_j)^T v_i \\
&= \sum_{i=1}^{k} {v_i}^T \Sigma_{\mathcal{X}} v_i
\end{aligned}$$

Captured variance

For any orthonormal $v_1, \ldots, v_k$

$$\sum_{i=1}^{k} \operatorname{var}(\mathcal{P}_{v_i}\,\mathcal{X}) = \sum_{i=1}^{k} \frac{1}{n} \sum_{j=1}^{n} {v_i}^T c(x_j) c(x_j)^T v_i$$

$$= \sum_{i=1}^{k} {v_i}^T \Sigma_{\mathcal{X}} v_i$$

By spectral theorem, eigenvectors optimize each individual term

Eigenvectors also optimize sum

For any symmetric $A \in \mathbb{R}^{d \times d}$ with eigenvectors $u_1, \ldots, u_k$

$$\sum_{i=1}^{k} u_i^T A u_i \geq \sum_{i=1}^{k} v_i^T A v_i.$$

for any k orthonormal vectors $v_1, \ldots, v_k$

Proof by induction on k

Base ($k = 1$)?

Proof by induction on k

Base ($k = 1$)? Follows from spectral theorem

Step

Let $\mathcal{S} := \text{span}(v_1, \ldots, v_k)$

For any orthonormal basis for $\mathcal{S}$ b_1, ..., b_k of $\mathcal{S}$

$$VV^T = BB^T$$

Step

Let $\mathcal{S} := \text{span}(v_1, \ldots, v_k)$

For any orthonormal basis for $\mathcal{S}$ $b_1, \ldots, b_k$ of $\mathcal{S}$

$$VV^T = BB^T$$

Choice of basis does not change cost function

$$\sum_{i=1}^{k} v_i^T A v_i$$

Step

Let $\mathcal{S} := \mathrm{span}(v_1, \ldots, v_k)$

For any orthonormal basis for $\mathcal{S}$ b_1, ..., b_k of $\mathcal{S}$

$$VV^T = BB^T$$

Choice of basis does not change cost function

$$\sum_{i=1}^{k} v_i^T A v_i = \mathrm{trace}\left(V^T A V\right)$$

Step

Let $\mathcal{S} := \text{span}(v_1, \ldots, v_k)$

For any orthonormal basis for $\mathcal{S}$ b_1, ..., b_k of $\mathcal{S}$

$$VV^T = BB^T$$

Choice of basis does not change cost function

$$\begin{aligned}\sum_{i=1}^{k} v_i^T A v_i &= \text{trace}\left(V^T A V\right) \\ &= \text{trace}\left(A V V^T\right)\end{aligned}$$

Step

Let $\mathcal{S} := \text{span}(v_1, \ldots, v_k)$

For any orthonormal basis for $\mathcal{S}$ $b_1, \ldots, b_k$ of $\mathcal{S}$

$$VV^T = BB^T$$

Choice of basis does not change cost function

$$\begin{aligned}
\textstyle\sum_{i=1}^{k} v_i^T A v_i &= \text{trace}\left(V^T A V\right) \\
&= \text{trace}\left(A V V^T\right) \\
&= \text{trace}\left(A B B^T\right)
\end{aligned}$$

Step

Let $\mathcal{S} := \text{span}(v_1, \ldots, v_k)$

For any orthonormal basis for $\mathcal{S}$ b_1, ..., b_k of $\mathcal{S}$

$$VV^T = BB^T$$

Choice of basis does not change cost function

$$\begin{aligned}
\sum_{i=1}^k v_i^T A v_i &= \text{trace}\left(V^T A V\right) \\
&= \text{trace}\left(A V V^T\right) \\
&= \text{trace}\left(A B B^T\right) \\
&= \sum_{i=1}^k b_i^T A b_i
\end{aligned}$$

Step

Let $\mathcal{S} := \text{span}(v_1, \ldots, v_k)$

For any orthonormal basis for $\mathcal{S}$ b_1, ..., b_k of $\mathcal{S}$

$$VV^T = BB^T$$

Choice of basis does not change cost function

$$\begin{aligned} \sum_{i=1}^{k} v_i^T A v_i &= \text{trace}\left(V^T A V\right) \\ &= \text{trace}\left(A V V^T\right) \\ &= \text{trace}\left(A B B^T\right) \\ &= \sum_{i=1}^{k} b_i^T A b_i \end{aligned}$$

Let's choose wisely

Step

We choose b orthogonal to $u_1, \ldots, u_{k-1}$

Step

We choose b orthogonal to $u_1, \ldots, u_{k-1}$

By spectral theorem

$$u_k^T A u_k \geq b^T A b$$

Step

We choose b orthogonal to $u_1, \ldots, u_{k-1}$

By spectral theorem

$$u_k^T A u_k \geq b^T A b$$

Now choose orthonormal basis $b_1, b_2, \ldots, b_k$ for $\mathcal{S}$ so that $b_k := b$

Step

We choose b orthogonal to $u_1, \ldots, u_{k-1}$

By spectral theorem

$$u_k^T A u_k \geq b^T A b$$

Now choose orthonormal basis $b_1, b_2, \ldots, b_k$ for $\mathcal{S}$ so that $b_k := b$

By induction assumption

$$\sum_{i=1}^{k-1} u_i^T A u_i \geq \sum_{i=1}^{k-1} b_i^T A b_i$$

Conclusion

For any k orthonormal vectors $v_1, \ldots, v_k$

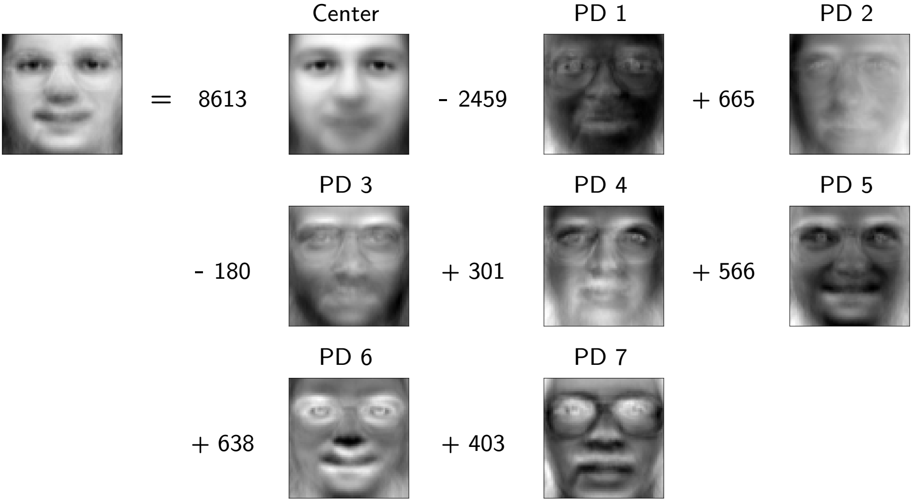
$$\sum_{i=1}^{k} \mathrm{var}(\mathrm{pc}[i]) \geq \sum_{i=1}^{k} \mathrm{var}(\mathcal{P}_{v_i}\,\mathcal{X}),$$

where $\mathrm{pc}[i] := \{\mathrm{pc}_1[i], \ldots, \mathrm{pc}_n[i]\} = \mathcal{P}_{u_i}\,\mathcal{X}$

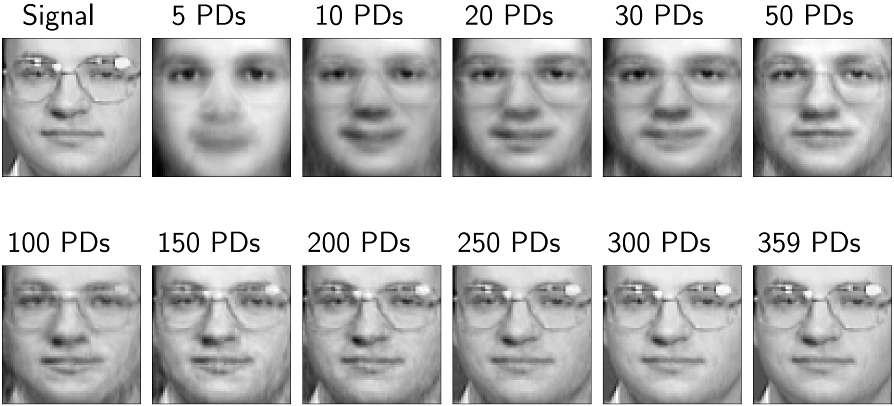
Faces

$$x_i^{\text{reduced}} := \text{av}(\mathcal{X}) + \sum_{j=1}^{7} \text{pc}_i[j] u_j$$

Projection onto first 7 principal directions



Projection onto first k principal directions



Nearest-neighbor classification

Training set of points and labels $\{x_1, l_1\}$, ..., $\{x_n, l_n\}$

To classify a new data point y, find

$$i^* := \arg \min_{1 \leq i \leq n} \|y - x_i\|_2 ,$$

and assign l_{i^*} to y

Cost: $\mathcal{O}(nd)$ to classify new point

Nearest neighbors in principal-component space

Idea: Project onto first k main principal directions beforehand

Costly reduced to $\mathcal{O}(kd)$

Computing eigendecomposition is costly, but only needs to be done once

Face recognition

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

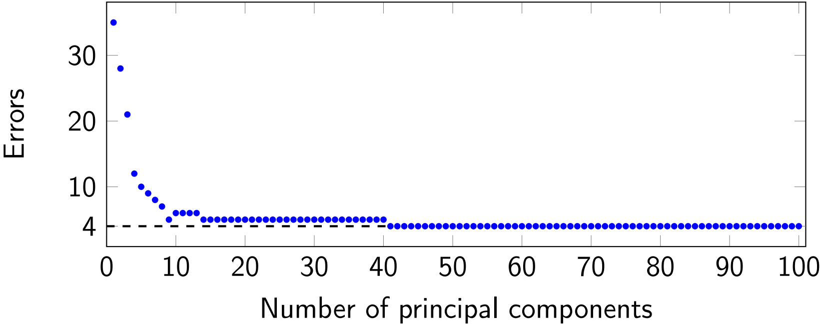
Test set: 1 new image from each subject

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

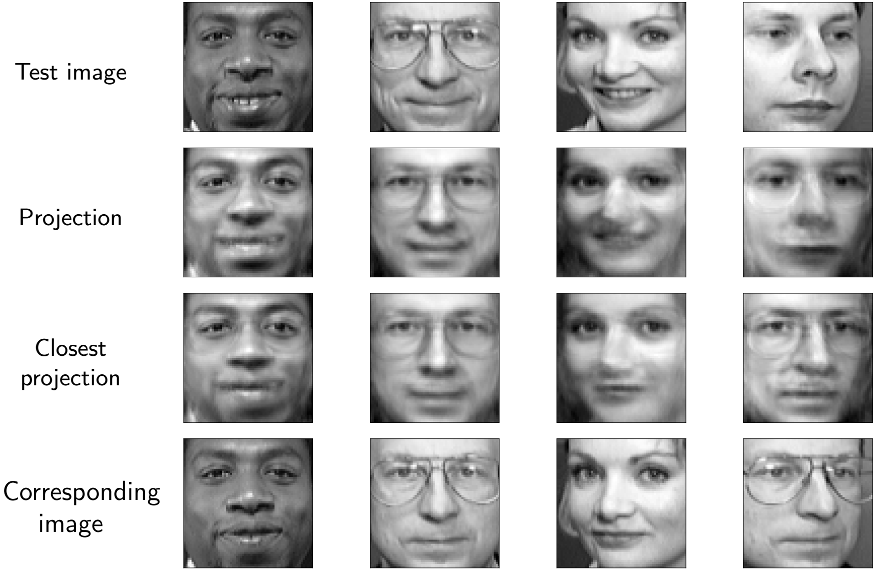
To classify we:

1. Project onto first k principal directions
2. Apply nearest-neighbor classification using the ℓ_2-norm distance in $\mathbb{R}^k$

Performance



Nearest neighbor in $\mathbb{R}^{41}$



Dimensionality reduction for visualization

Motivation: Visualize high-dimensional features projected onto 2D or 3D

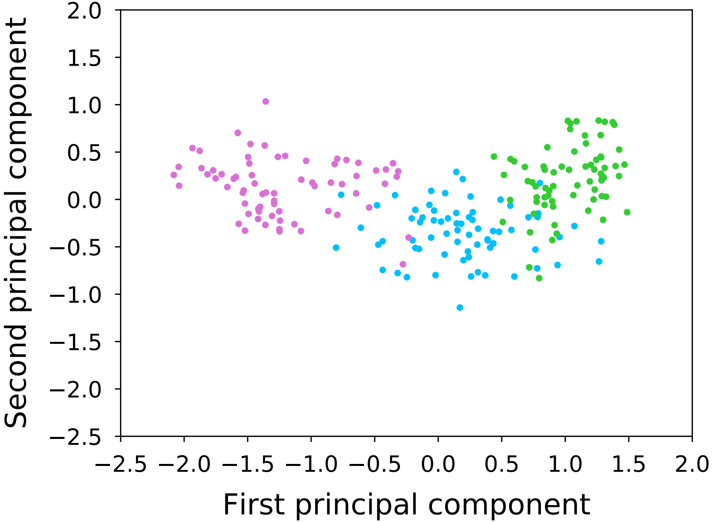
Example:

Seeds from three different varieties of wheat: Kama, Rosa and Canadian

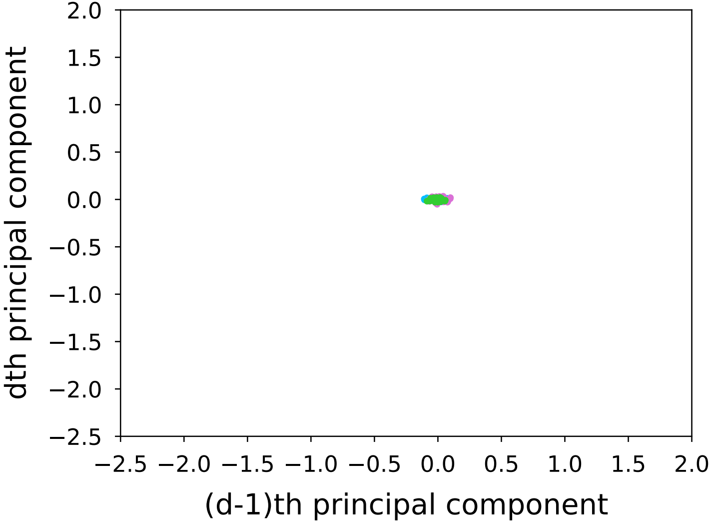
Features:

- Area
- Perimeter
- Compactness
- Length of kernel
- Width of kernel
- Asymmetry coefficient
- Length of kernel groove

Projection onto two first PDs



Projection onto two last PDs



Covariance matrix

The spectral theorem

Principal component analysis

Dimensionality reduction via PCA

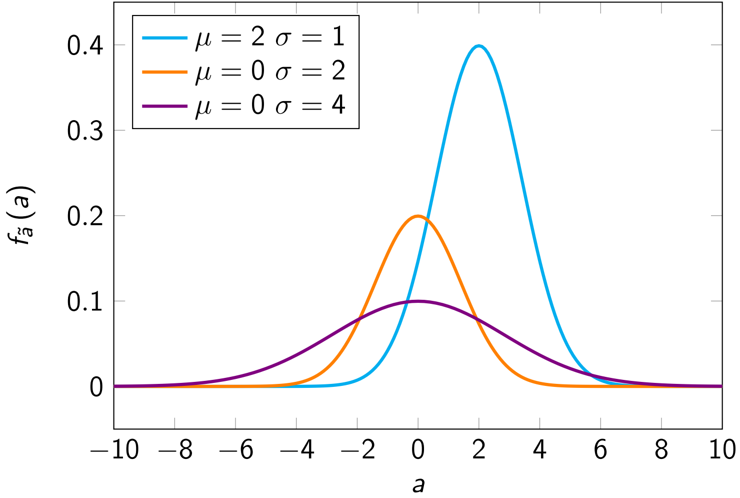
Gaussian random vectors

Gaussian random variables

The pdf of a Gaussian or normal random variable $\tilde{a}$ with mean μ and standard deviation σ is given by

$$f_{\tilde{a}}(a) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(a-\mu)^2}{2\sigma^2}}$$

Gaussian random variables



Gaussian random variables

$$\mu = \int_{a=-\infty}^{\infty} a f_{\tilde{a}}(a) \, \mathrm{d}a$$

$$\sigma^2 = \int_{a=-\infty}^{\infty} (a - \mu)^2 f_{\tilde{a}}(a) \, \mathrm{d}a$$

Linear transformation of Gaussian

If $\tilde{a}$ is a Gaussian random variable with mean μ and standard deviation σ, then for any $\alpha, \beta \in \mathbb{R}$

$$\tilde{b} := \alpha \tilde{a} + \beta$$

is a Gaussian random variable with $\alpha\mu + \beta$ and standard deviation $|\alpha| \sigma$

Proof

Let $\alpha > 0$ (proof for $a < 0$ is very similar),

$F_{\tilde{b}}(b) =$

Proof

Let $\alpha > 0$ (proof for $a < 0$ is very similar),

$$F_{\tilde{b}}(b) = \mathrm{P}\left(\tilde{b} \leq b\right)$$

Proof

Let $\alpha > 0$ (proof for $a < 0$ is very similar),

$$\begin{aligned} F_{\tilde{b}}(b) &= \mathrm{P}\left(\tilde{b} \leq b\right) \\ &= \mathrm{P}\left(\alpha\tilde{b} + \beta \leq b\right) \end{aligned}$$

Proof

Let $\alpha > 0$ (proof for $a < 0$ is very similar),

$$\begin{aligned} F_{\tilde{b}}(b) &= \mathrm{P}\left(\tilde{b} \leq b\right) \\ &= \mathrm{P}\left(\alpha\tilde{b} + \beta \leq b\right) \\ &= \mathrm{P}\left(\tilde{b} \leq \frac{b - \beta}{\alpha}\right) \end{aligned}$$

Proof

Let $\alpha > 0$ (proof for $a < 0$ is very similar),

$$
\begin{aligned}
F_{\tilde{b}}(b) &= \mathrm{P}\left(\tilde{b} \leq b\right) \\
&= \mathrm{P}\left(\alpha\tilde{b} + \beta \leq b\right) \\
&= \mathrm{P}\left(\tilde{b} \leq \frac{b-\beta}{\alpha}\right) \\
&= \int_{-\infty}^{\frac{b-\beta}{\alpha}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(a-\mu)^2}{2\sigma^2}} \, \mathrm{d}a
\end{aligned}
$$

Proof

Let $\alpha > 0$ (proof for $a < 0$ is very similar),

$$
\begin{aligned}
F_{\tilde{b}}(b) &= \mathrm{P}\left(\tilde{b} \leq b\right) \\
&= \mathrm{P}\left(\alpha\tilde{b} + \beta \leq b\right) \\
&= \mathrm{P}\left(\tilde{b} \leq \frac{b-\beta}{\alpha}\right) \\
&= \int_{-\infty}^{\frac{b-\beta}{\alpha}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(a-\mu)^2}{2\sigma^2}} \, \mathrm{d}a \\
&= \int_{-\infty}^{b} \frac{1}{\sqrt{2\pi}\alpha\sigma} e^{-\frac{(w-\alpha\mu-\beta)^2}{2\alpha^2\sigma^2}} \, \mathrm{d}w \qquad \text{change of variables } w := \alpha a + \beta
\end{aligned}
$$

Proof

Let $\alpha > 0$ (proof for $a < 0$ is very similar),

$$
\begin{aligned}
F_{\tilde{b}}(b) &= \mathrm{P}\left(\tilde{b} \leq b\right) \\
&= \mathrm{P}\left(\alpha \tilde{b} + \beta \leq b\right) \\
&= \mathrm{P}\left(\tilde{b} \leq \frac{b-\beta}{\alpha}\right) \\
&= \int_{-\infty}^{\frac{b-\beta}{\alpha}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(a-\mu)^2}{2\sigma^2}} \, \mathrm{d}a \\
&= \int_{-\infty}^{b} \frac{1}{\sqrt{2\pi}\alpha\sigma} e^{-\frac{(w-\alpha\mu-\beta)^2}{2\alpha^2\sigma^2}} \, \mathrm{d}w \qquad \text{change of variables } w := \alpha a + \beta
\end{aligned}
$$

Differentiating with respect to b:

$$
f_{\tilde{b}}(b) = \frac{1}{\sqrt{2\pi}\alpha\sigma} e^{-\frac{(b-\alpha\mu-\beta)^2}{2\alpha^2\sigma^2}}
$$

Gaussian random vector

A Gaussian random vector $\tilde{x}$ is a random vector with joint pdf

$$f_{\tilde{x}}(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

where $\mu \in \mathbb{R}^d$ is the mean and $\Sigma \in \mathbb{R}^{d \times d}$ the covariance matrix

$\Sigma \in \mathbb{R}^{d \times d}$ is positive definite (positive eigenvalues)

Contour surfaces

Set of points at which pdf is constant

$$c = x^T \Sigma^{-1} x \qquad \text{assuming } \mu = 0$$

Contour surfaces

Set of points at which pdf is constant

$$
\begin{aligned}
c &= x^T \Sigma^{-1} x \qquad \text{assuming } \mu = 0 \\
&= x^T U \Lambda^{-1} U x
\end{aligned}
$$

Contour surfaces

Set of points at which pdf is constant

$$
\begin{aligned}
c &= x^T \Sigma^{-1} x \qquad \text{assuming } \mu = 0 \\
&= x^T U \Lambda^{-1} U x \\
&= \sum_{i=1}^{d} \frac{(u_i^T x)^2}{\sqrt{\lambda_i}}
\end{aligned}
$$

Contour surfaces

Set of points at which pdf is constant

$$
\begin{aligned}
c &= x^T \Sigma^{-1} x \qquad \text{assuming } \mu = 0 \\
&= x^T U \Lambda^{-1} U x \\
&= \sum_{i=1}^{d} \frac{(u_i^T x)^2}{\sqrt{\lambda_i}}
\end{aligned}
$$

Ellipsoid with axes proportional to $\sqrt{\lambda_i}$

2D example

$$\mu = 0$$

$$\Sigma = \begin{bmatrix} 0.5 & -0.3 \\ -0.3 & 0.5 \end{bmatrix}$$

2D example

$$\mu = 0$$

$$\Sigma = \begin{bmatrix} 0.5 & -0.3 \\ -0.3 & 0.5 \end{bmatrix}$$

$$\lambda_1 = 0.8$$

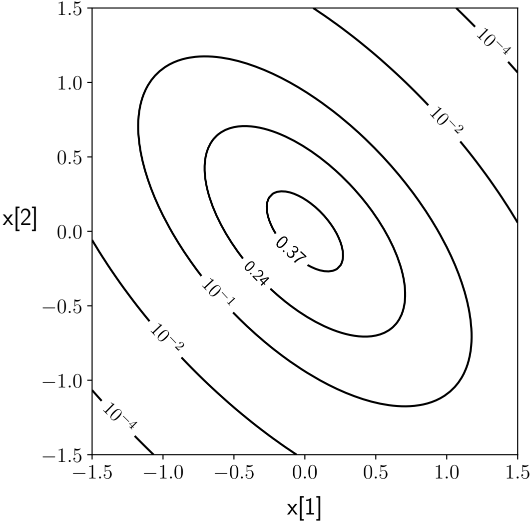
$$\lambda_2 = 0.2$$

$$u_1 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

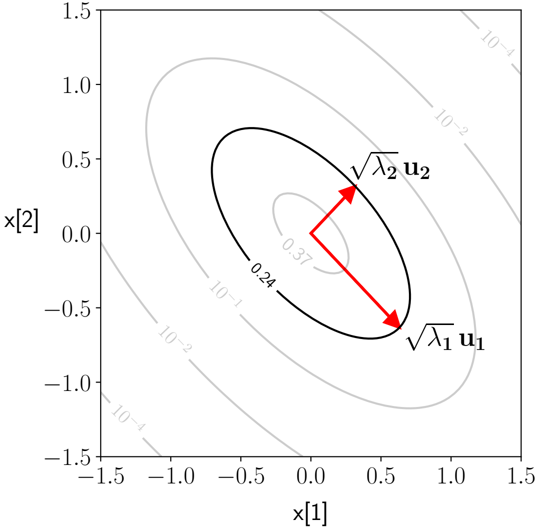
$$u_2 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

How does the ellipse look like?

Contour surfaces



Contour surfaces



Uncorrelation implies independence

If the covariance matrix is diagonal,

$$\Sigma_{\tilde{x}} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix}$$

the entries of a Gaussian random vector are independent

Proof

$$\Sigma_{\tilde{x}}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_d^2} \end{bmatrix}$$

$$|\Sigma| = \prod_{i=1}^{d} \sigma_i^2$$

Proof

$$f_{\tilde{x}}(x)$$

Proof

$$f_{\tilde{x}}(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

Proof

$$
\begin{aligned}
f_{\tilde{x}}\left(x\right) &= \frac{1}{\sqrt{\left(2\pi\right)^{d}\left|\Sigma\right|}} \exp\left(-\frac{1}{2}\left(x-\mu\right)^{T}\Sigma^{-1}\left(x-\mu\right)\right) \\
&= \prod_{i=1}^{d} \frac{1}{\sqrt{(2\pi)}\sigma_i} \exp\left(-\frac{\left(x_i-\mu_i\right)^{2}}{2\sigma_i^{2}}\right)
\end{aligned}
$$

Proof

$$
\begin{aligned}
f_{\tilde{x}}\left(x\right) &= \frac{1}{\sqrt{\left(2\pi\right)^{d}\left|\Sigma\right|}}\exp\left(-\frac{1}{2}\left(x-\mu\right)^{T}\Sigma^{-1}\left(x-\mu\right)\right) \\
&= \prod_{i=1}^{d}\frac{1}{\sqrt{(2\pi)}\sigma_i}\exp\left(-\frac{\left(x_i-\mu_i\right)^2}{2\sigma_i^2}\right) \\
&= \prod_{i=1}^{d} f_{\tilde{x}_i}\left(x_i\right)
\end{aligned}
$$

Linear transformations

Let $\tilde{x}$ be a Gaussian random vector of dimension d with mean μ and covariance matrix Σ

For any matrix $A \in \mathbb{R}^{m \times d}$ and $\vec{b} \in \mathbb{R}^m$ $\tilde{y} = A\tilde{x} + \vec{b}$ is Gaussian with mean $A\mu + \vec{b}$ and covariance matrix $A\Sigma A^T$ (as long as it is full rank)

PCA on Gaussian random vectors

Let $\tilde{x}$ be a Gaussian random vector with covariance matrix $\Sigma := U \Lambda U^T$

The principal components

$$\widetilde{pc} := U^T \tilde{x}$$

are Gaussian and have covariance matrix

$$U^T \Sigma U = \Lambda$$

PCA on Gaussian random vectors

Let $\tilde{x}$ be a Gaussian random vector with covariance matrix $\Sigma := U \Lambda U^T$

The principal components

$$\widetilde{pc} := U^T \tilde{x}$$

are Gaussian and have covariance matrix

$$U^T \Sigma U = \Lambda$$

so they are independent

PCA on Gaussian random vectors

Let $\tilde{x}$ be a Gaussian random vector with covariance matrix $\Sigma := U \Lambda U^T$

The principal components

$$\widetilde{pc} := U^T \tilde{x}$$

are Gaussian and have covariance matrix

$$U^T \Sigma U = \Lambda$$

so they are independent

Often not the case in practice!

Maximum likelihood for Gaussian vectors

Log-likelihood of Gaussian parameters

$$
\begin{aligned}
&(\mu_{\text{ML}}, \Sigma_{\text{ML}}) \\
&:= \arg \max_{\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}} \log \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)\right) \\
&= \arg \min_{\mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d \times d}} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu).
\end{aligned}
$$

Solution is sample mean and variance

Additional justification, but PCA is useful without Gaussian assumption!