Principal Component Analysis

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Motivation

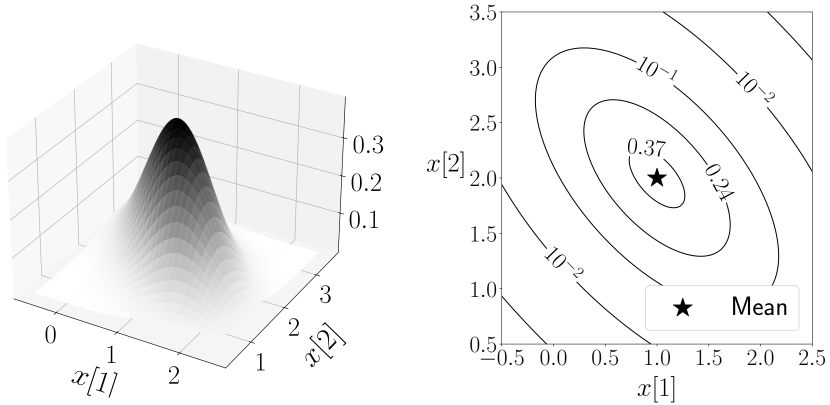
Describe data with multiple features

Model: d-dimensional random vector

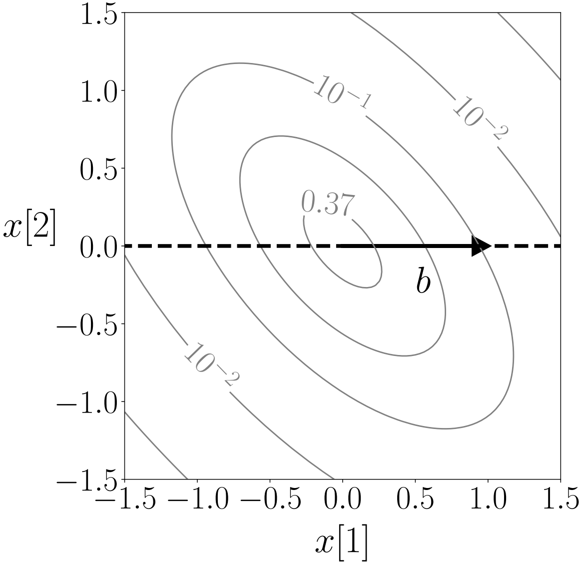
$$\tilde{x} := \begin{bmatrix} \tilde{x}[1] \\ \tilde{x}[2] \\ \cdots \\ \tilde{x}[d] \end{bmatrix}$$

Idea: Identify directions in which $\tilde{x}$ has more variance

Gaussian random vector



Variance in a certain direction?



Variance in a certain direction?

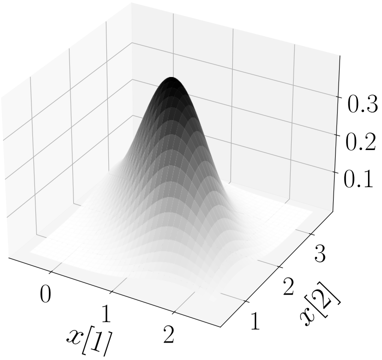
After centering by subtracting the mean, if $||b||_2 = 1$

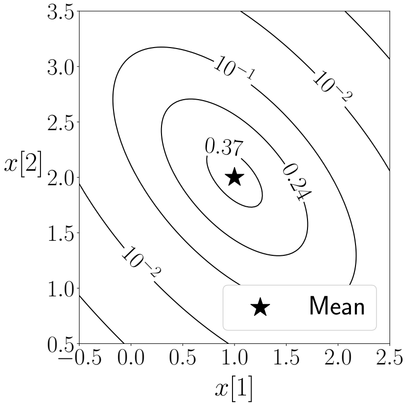
$$\tilde{x} = \underbrace{(b^T\tilde{x})b}_{\text{collinear with } b} + \underbrace{\tilde{x} - (b^T\tilde{x})b}_{\text{orthogonal to } b}$$

Variance of a linear combination

$$\mathrm{Var}\left[a^T \tilde{x}\right] = a^T \Sigma_{\tilde{x}} a$$

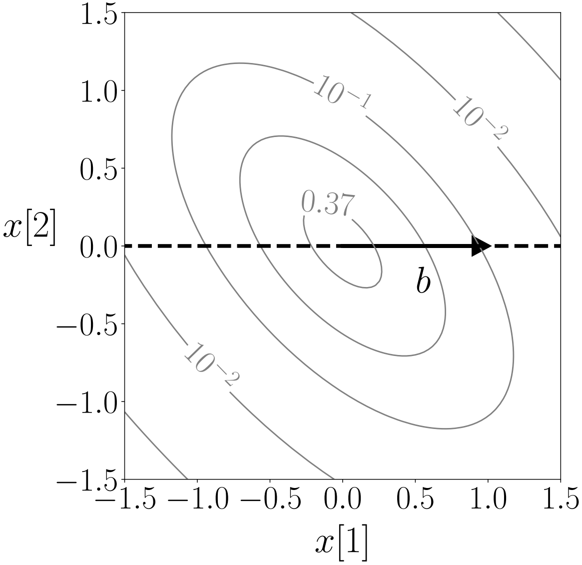
Gaussian random vector





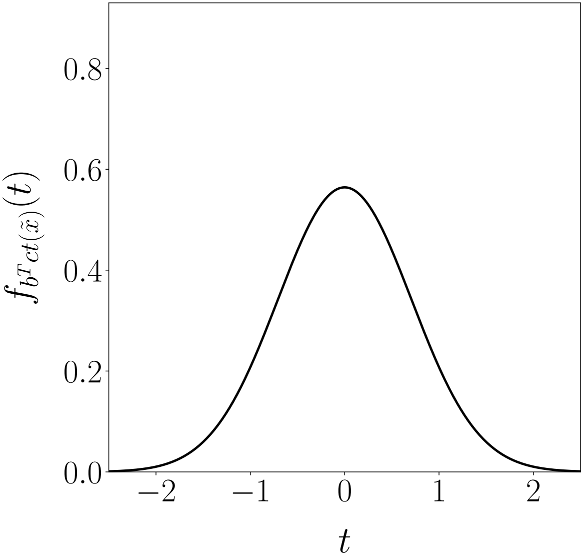
$$\Sigma_{\tilde{x}} := \begin{bmatrix} 0.5 & -0.3 \\ -0.3 & 0.5 \end{bmatrix}$$

Variance in a certain direction?

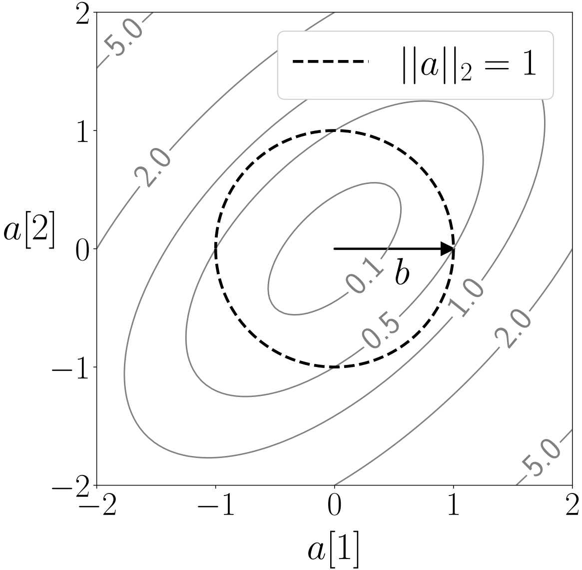


Variance in a certain direction

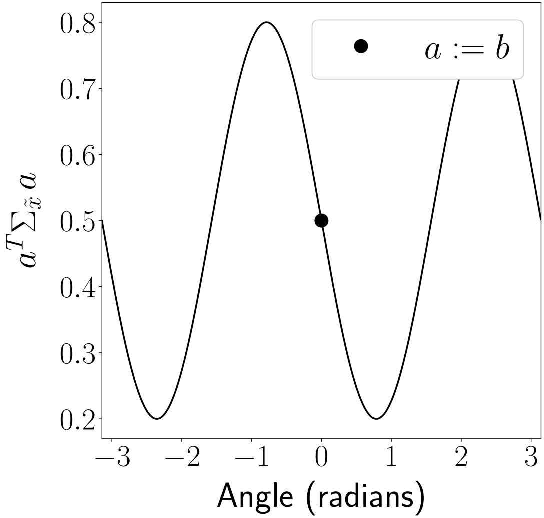
$$\mathrm{Var}[b^T \tilde{x}] = b^T \Sigma_{\tilde{x}} b = 0.5$$



Quadratic form $a^T \Sigma_{\tilde{x}} a = \mathrm{Var}[a^T \tilde{x}]$



$a^T \Sigma_{\tilde{x}} a$ on the unit circle



Maximum is direction of maximum variance

Spectral theorem

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

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

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

$$\lambda_1 = \max_{||a||_2=1} a^T M a$$

$$u_1 = \arg\max_{||a||_2=1} a^T M a$$

$$\lambda_k = \max_{||a||_2=1, a\perp u_1,\ldots,u_{k-1}} a^T M a, \quad 2 \leq k \leq d-1$$

$$u_k = \arg\max_{||a||_2=1, a\perp u_1,\ldots,u_{k-1}} a^T M a, \quad 2 \leq k \leq d-1$$

$$\lambda_d = \min_{||a||_2=1} a^T M a$$

$$u_d = \arg\min_{||a||_2=1} a^T M a$$

Is the covariance matrix symmetric?

$$\begin{aligned}
\Sigma_{\tilde{x}}^T &= \left(\mathrm{E}\left[\tilde{x}\tilde{x}^T\right]\right)^T \\
&= \mathrm{E}\left[\left(\tilde{x}\tilde{x}^T\right)^T\right] \\
&= \mathrm{E}\left[\tilde{x}\tilde{x}^T\right] = \Sigma_{\tilde{x}}
\end{aligned}$$

Principal directions

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

$$\lambda_1 = \max_{||a||_2=1} a^T \Sigma_{\tilde{x}} a = \max_{||a||_2=1} \mathrm{Var}[a^T \tilde{x}]$$

$$u_1 = \arg \max_{||a||_2=1} \mathrm{Var}[a^T \tilde{x}]$$

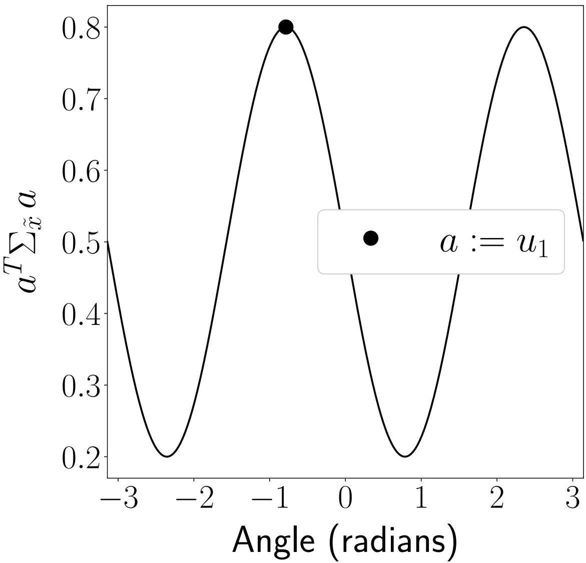
$$\lambda_k = \max_{||a||_2=1, a \perp u_1, \ldots, u_{k-1}} \mathrm{Var}[a^T \tilde{x}], \quad 2 \leq k \leq d$$

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

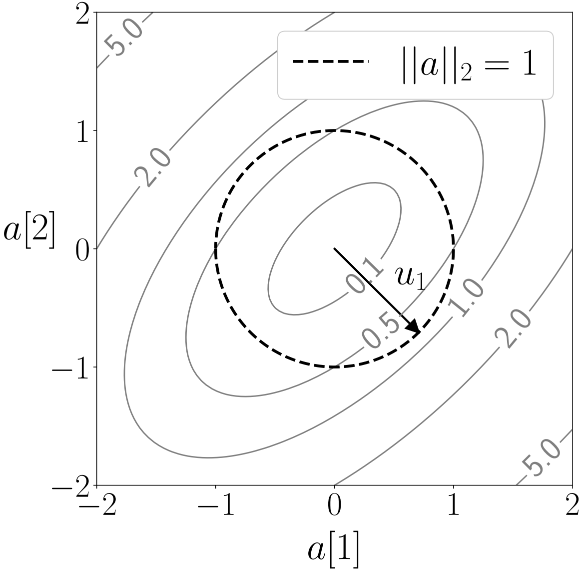
$$\lambda_d = \min_{||a||_2=1} \mathrm{Var}[a^T \tilde{x}]$$

$$u_d = \arg \min_{||a||_2=1} \mathrm{Var}[a^T \tilde{x}]$$

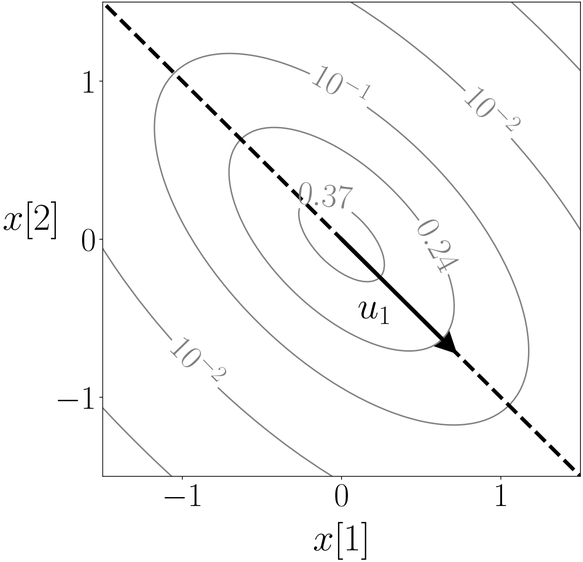
First principal direction



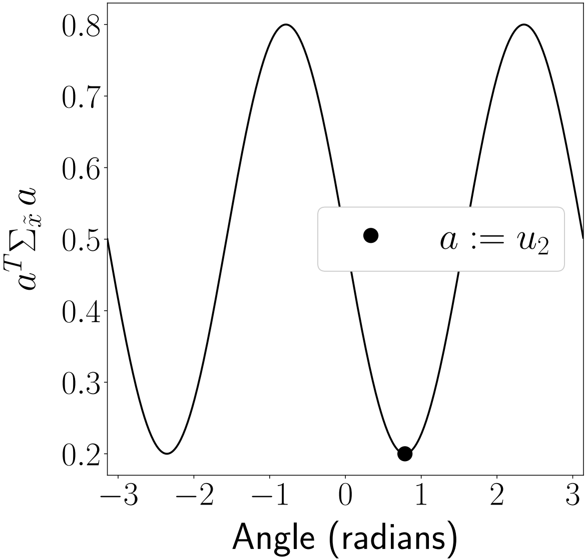
Quadratic form $a^T \Sigma_{\tilde{x}} a = \mathrm{Var}[a^T \tilde{x}]$



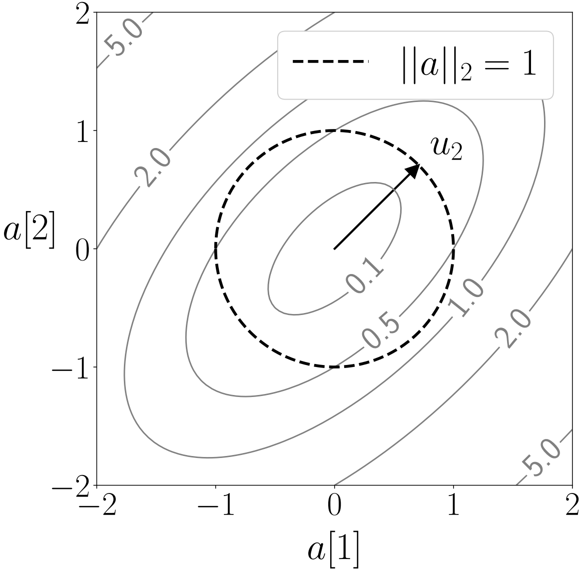
Joint pdf of $\tilde{x}$



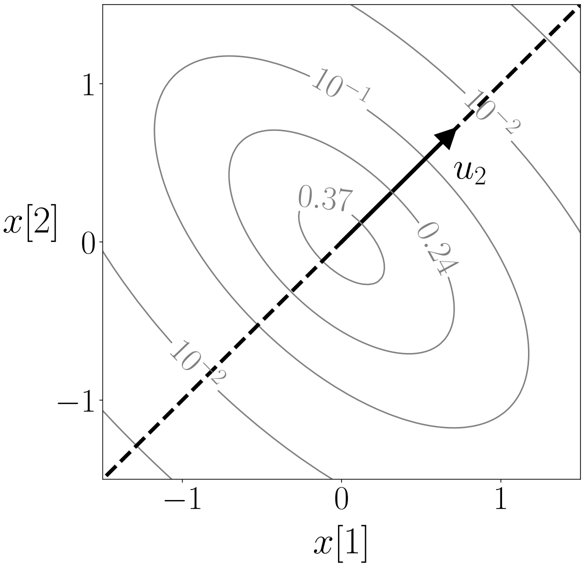
Second principal direction



Quadratic form $a^T \Sigma_{\tilde{x}} a = \mathrm{Var}[a^T \tilde{x}]$



Joint pdf of $\tilde{x}$



Principal components

Let $\text{ct}(\tilde{x}) := \tilde{x} - \mathrm{E}[\tilde{x}]$

$$\widetilde{w}_i := u_i^T \, \text{ct}(\tilde{x}) \qquad 1 \leq i \leq d$$

is the ith principal component

Variance of principal components

$$\begin{aligned}\operatorname{Var}\left[\widetilde{w}_i\right] &= u_i^T \Sigma_{\tilde{x}} u_i \\ &= \lambda_i u_i^T u_i \\ &= \lambda_i\end{aligned}$$

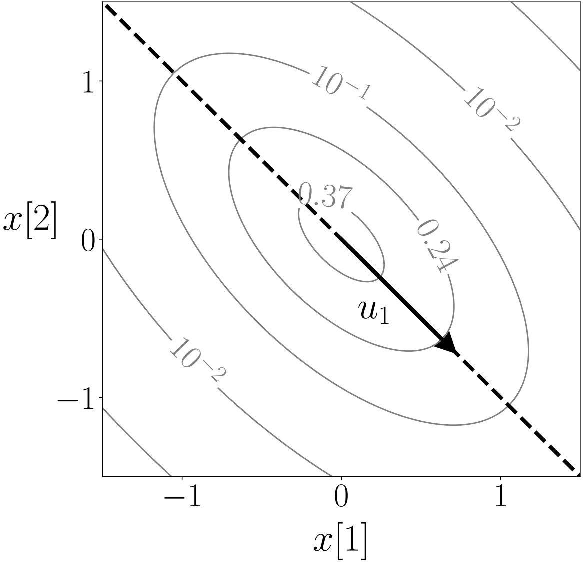
Spectral theorem:

$$\lambda_1 = \max_{||a||_2=1} \operatorname{Var}[a^T \tilde{x}]$$

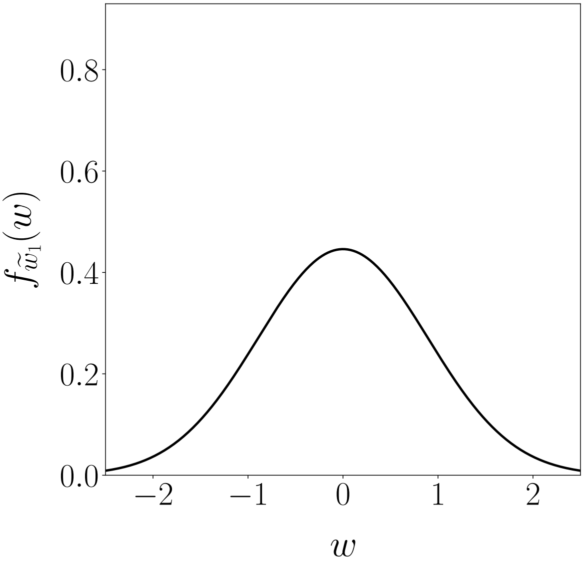
$$\lambda_k = \max_{||a||_2=1, a \perp u_1, \ldots, u_{k-1}} \operatorname{Var}[a^T \tilde{x}], \quad 2 \leq k \leq d$$

$$\lambda_d = \min_{||a||_2=1} \operatorname{Var}[a^T \tilde{x}]$$

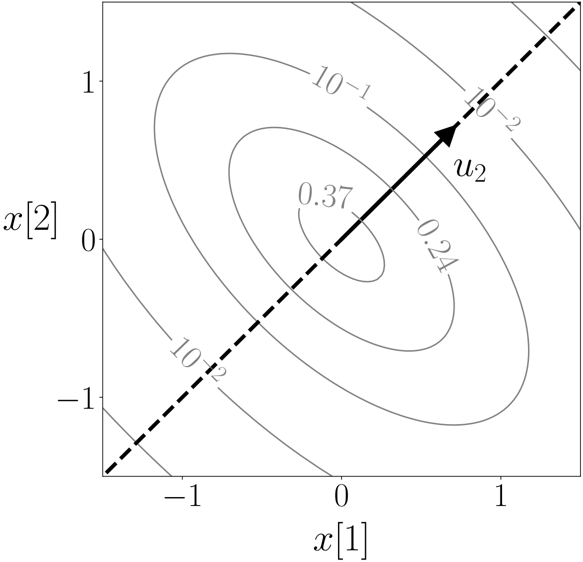
First principal direction



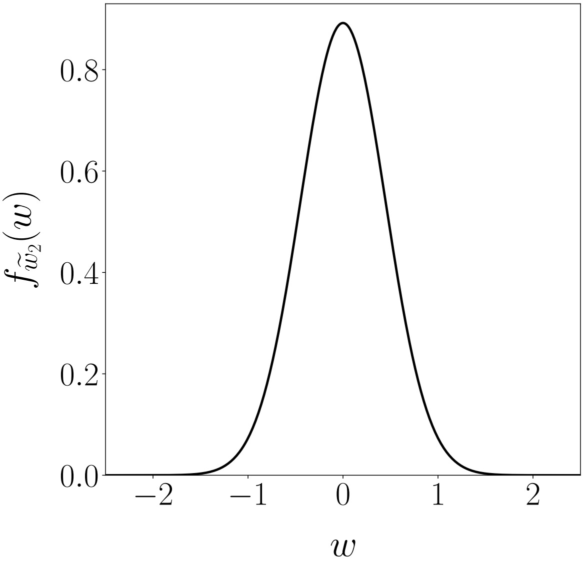
First principal component



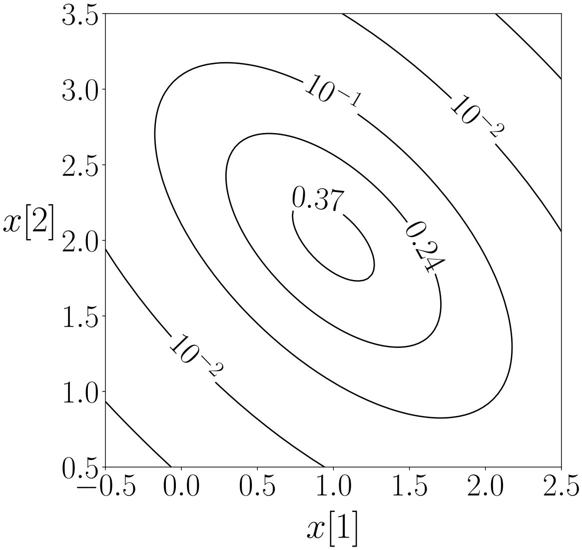
Second principal direction



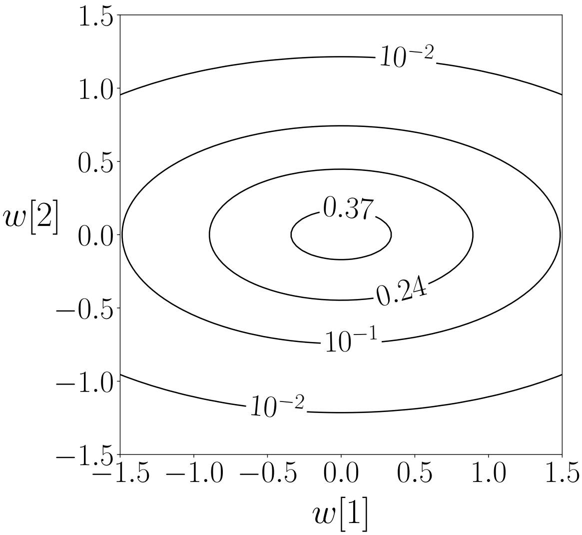
Second principal component



Original joint pdf



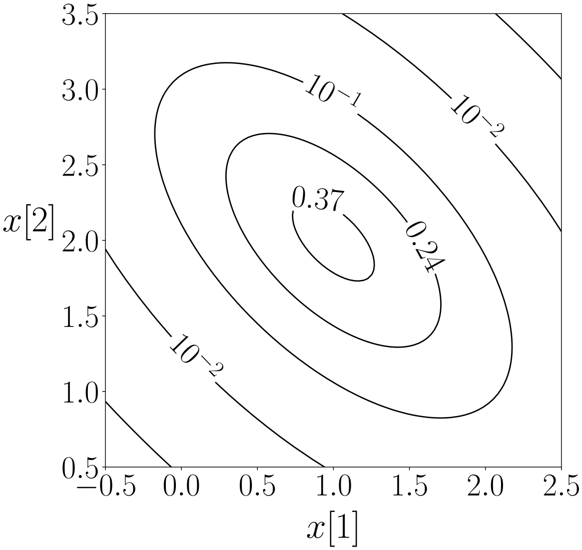
Joint pdf of principal components



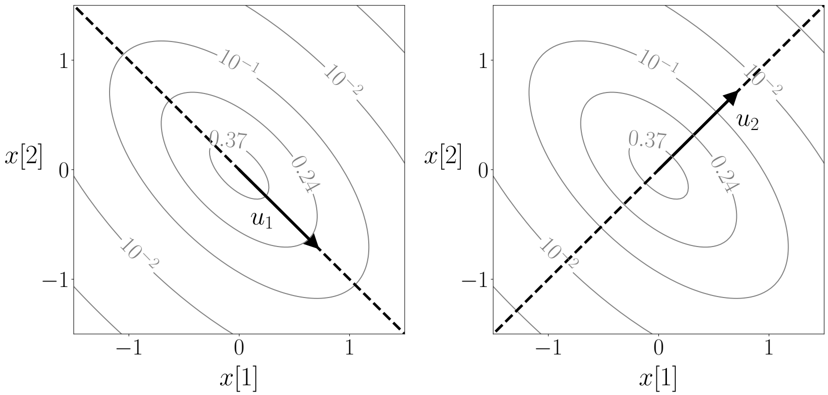
Correlation

$$\begin{aligned}
\mathrm{E}[\widetilde{w}_i \widetilde{w}_j] &= \mathrm{E}\left[u_i^T \operatorname{ct}(\tilde{x}) \operatorname{ct}(\tilde{x})^T u_j\right] \\
&= u_i^T \mathrm{E}[\operatorname{ct}(\tilde{x}) \operatorname{ct}(\tilde{x})^T] u_j \\
&= u_i^T \Sigma_{\tilde{x}} u_j \\
&= \lambda_j u_i^T u_j \\
&= 0
\end{aligned}$$

Gaussian random vector



Principal directions



PCA of data

Dataset $X = \{x_1, x_2, \ldots, x_n\}$ with d features

1. Compute sample covariance matrix Σ_X

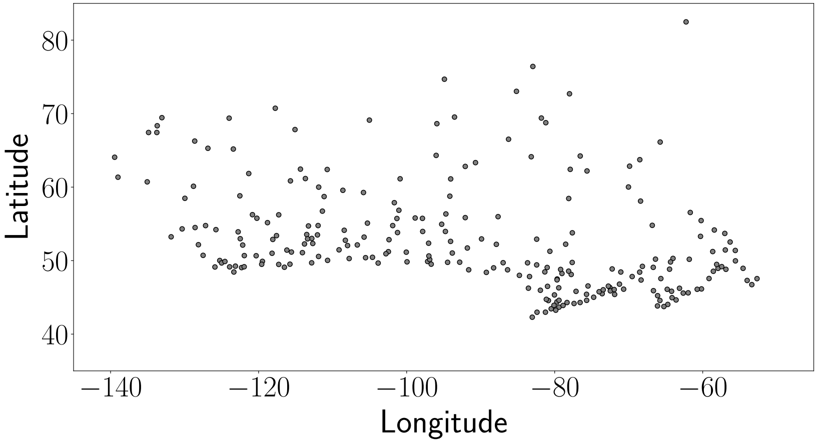
2. Eigendecomposition of Σ_X yields principal directions $u_1, \ldots, u_d$

3. Center the data and compute principal components

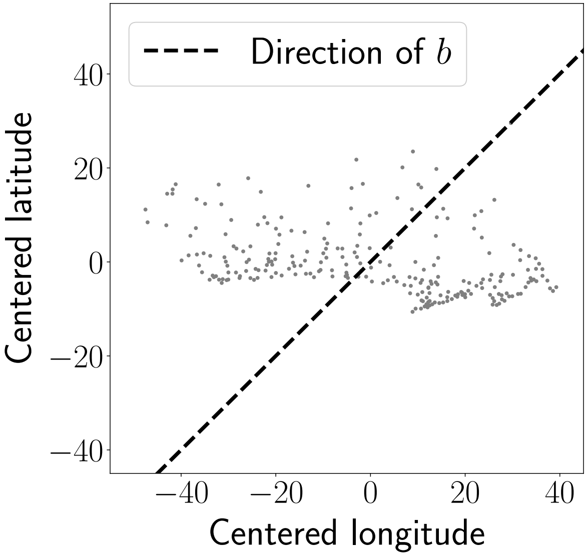
$$w_j[i] := u_j^T \operatorname{ct}(x_i), \quad 1 \leq i \leq n, \ 1 \leq j \leq d$$

where $\operatorname{ct}(x_i) := x_i - m(X)$

Cities in Canada



Variance in a certain direction?



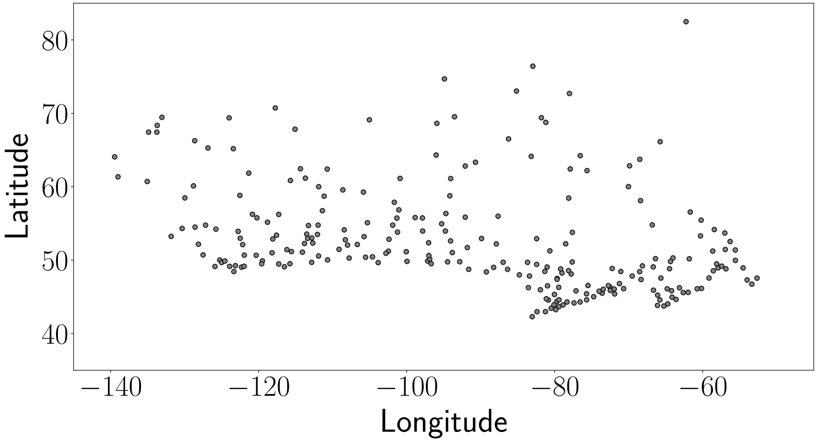
Sample variance of linear combination

Dataset: $X = \{x_1, \ldots, x_n\}$

$$X_a := \left\{ a^T x_1, \ldots, a^T x_n \right\}$$

$$v(X_a) = a^T \Sigma_X a$$

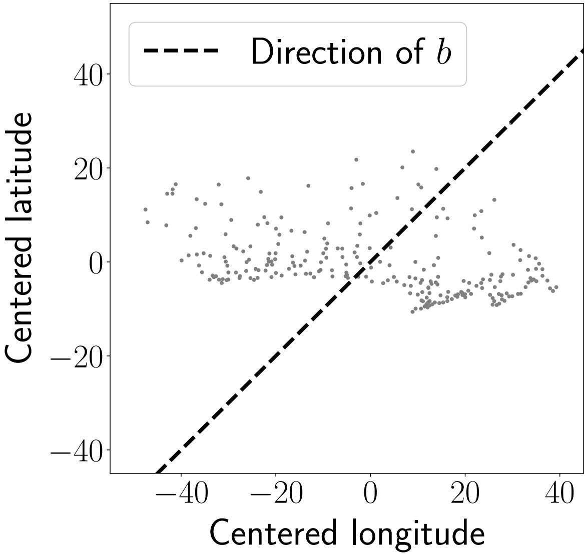
Cities in Canada



Sample covariance matrix:

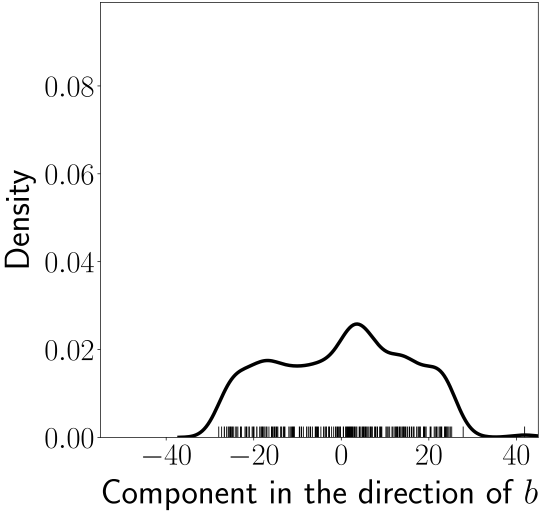
$$\Sigma_X = \begin{bmatrix} 524.9 & -59.8 \\ -59.8 & 53.7 \end{bmatrix}$$

Variance in a certain direction?

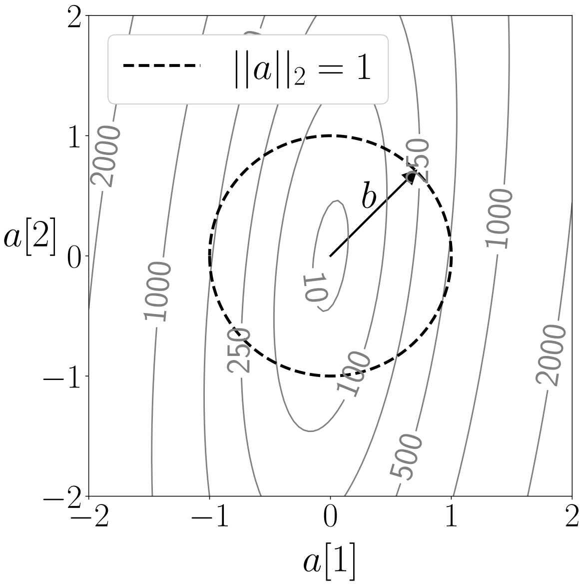


Variance in a certain direction

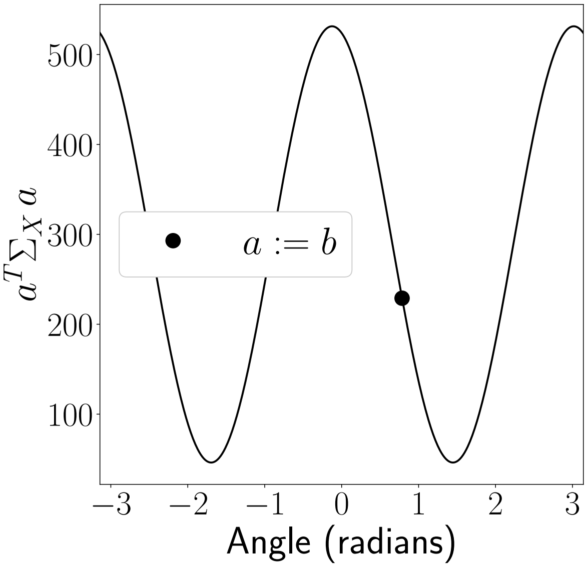
$$v(X_b) = b^T \Sigma_X b = 229$$



Quadratic form $a^T \Sigma_X a = v(X_a)$



$a^T \Sigma_X a$ on the unit circle



Spectral theorem

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

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

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

$$\lambda_1 = \max_{||a||_2=1} a^T M a$$

$$u_1 = \arg \max_{||a||_2=1} a^T M a$$

$$\lambda_k = \max_{||a||_2=1, a \perp u_1, \ldots, u_{k-1}} a^T M a, \quad 2 \leq k \leq d-1$$

$$u_k = \arg \max_{||a||_2=1, a \perp u_1, \ldots, u_{k-1}} a^T M a, \quad 2 \leq k \leq d-1$$

$$\lambda_d = \min_{||a||_2=1} a^T M a$$

$$u_d = \arg \min_{||a||_2=1} a^T M a$$

Principal directions

Let $u_1, \ldots, u_d$ be the eigenvectors, and $\lambda_1 > \ldots > \lambda_d$ the eigenvalues of Σ_X

$$\lambda_1 = \max_{||a||_2=1} a^T \Sigma_X a = \max_{||a||_2=1} v(X_a)$$

$$u_1 = \arg\max_{||a||_2=1} v(X_a)$$

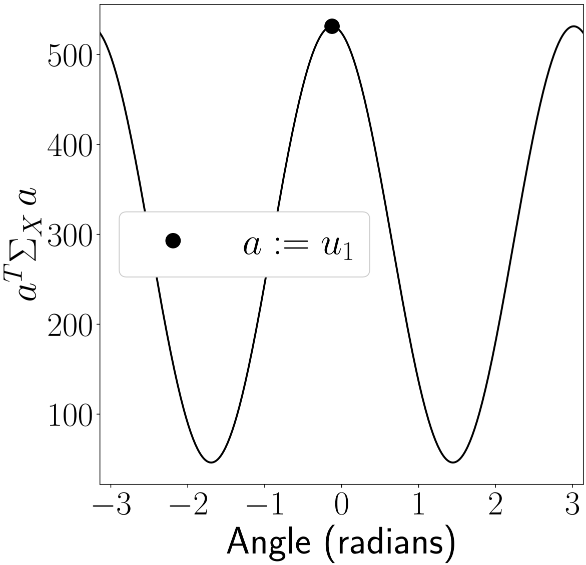
$$\lambda_k = \max_{||a||_2=1, a \perp u_1, \ldots, u_{k-1}} v(X_a), \quad 2 \leq k \leq d$$

$$u_k = \arg\max_{||a||_2=1, a \perp u_1, \ldots, u_{k-1}} v(X_a), \quad 2 \leq k \leq d$$

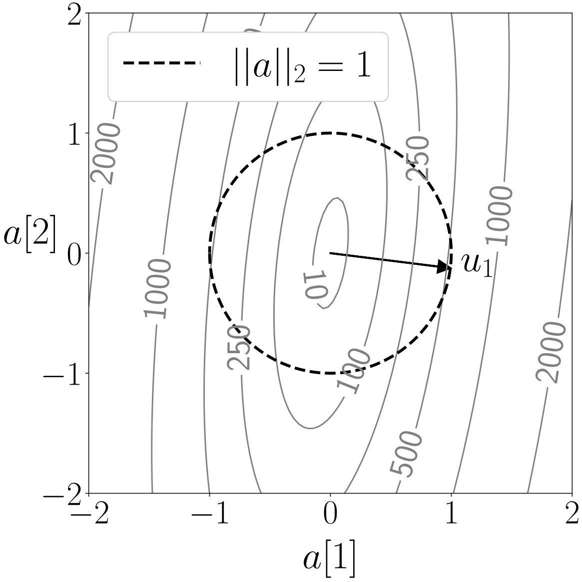
$$\lambda_d = \min_{||a||_2=1} v(X_a)$$

$$u_d = \arg\min_{||a||_2=1} v(X_a)$$

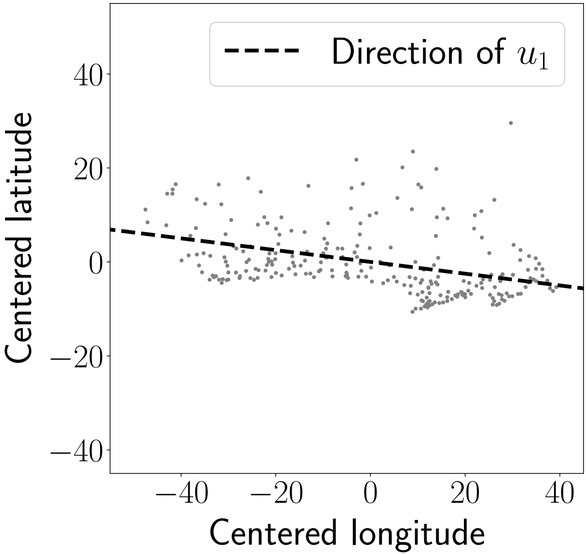
First principal direction



Quadratic form $a^T \Sigma_X a = v(X_a)$

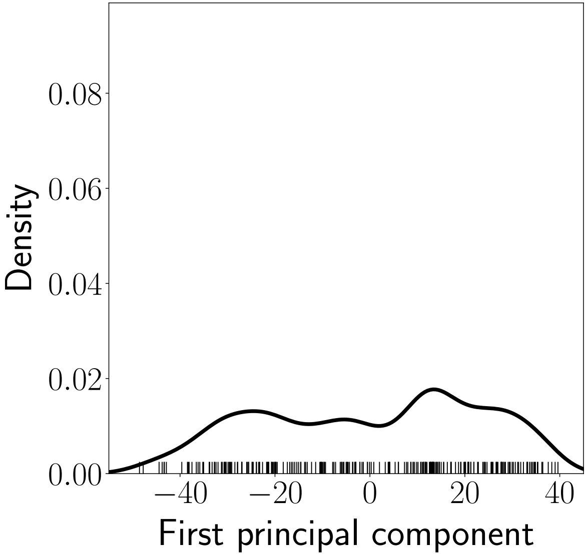


Data

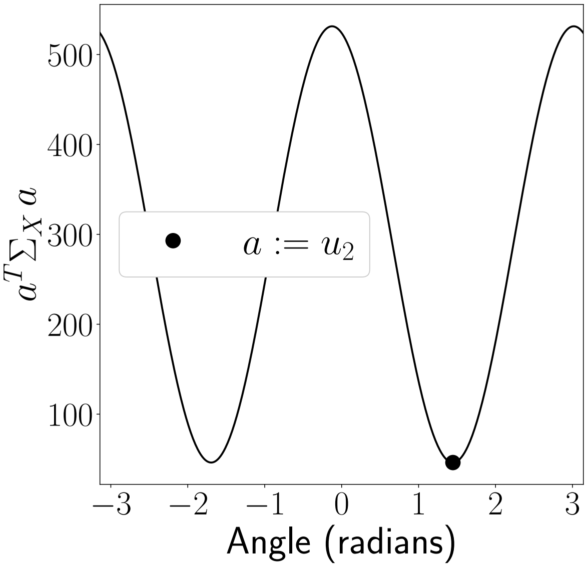


Direction of u_1
Centered latitude
Centered longitude
40
20
0
−20
−40
−40
−20
0
20
40

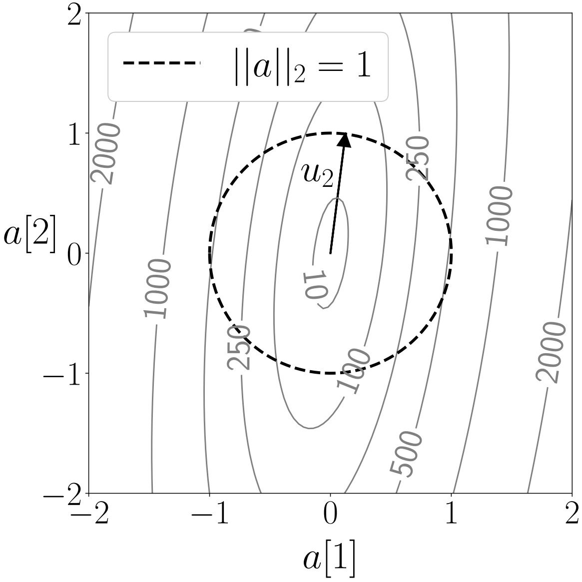
First principal component



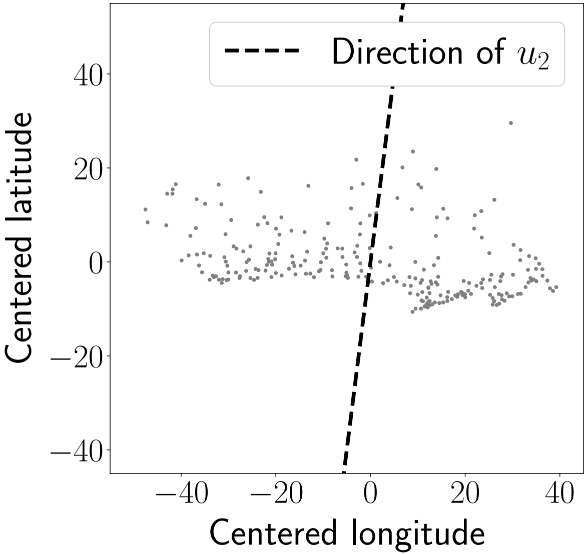
Second principal direction



Quadratic form $a^T \Sigma_X a = v(X_a)$

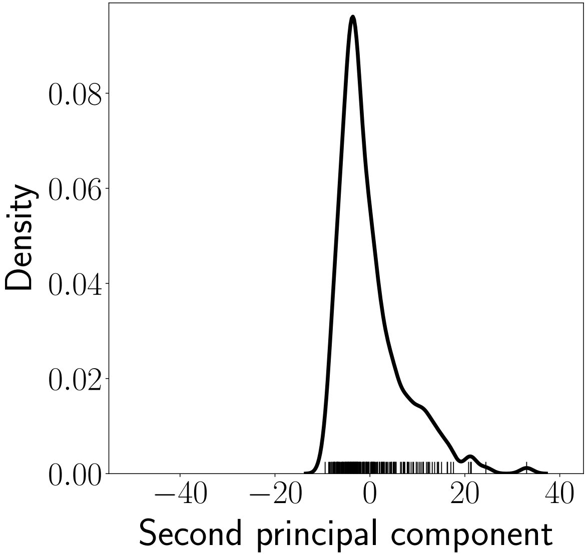


Data

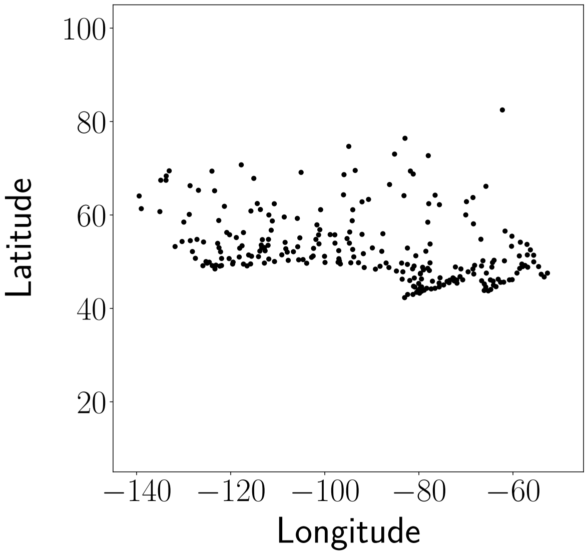


Direction of u_2
Centered latitude
Centered longitude
40
20
0
−20
−40
−40
−20
0
20
40

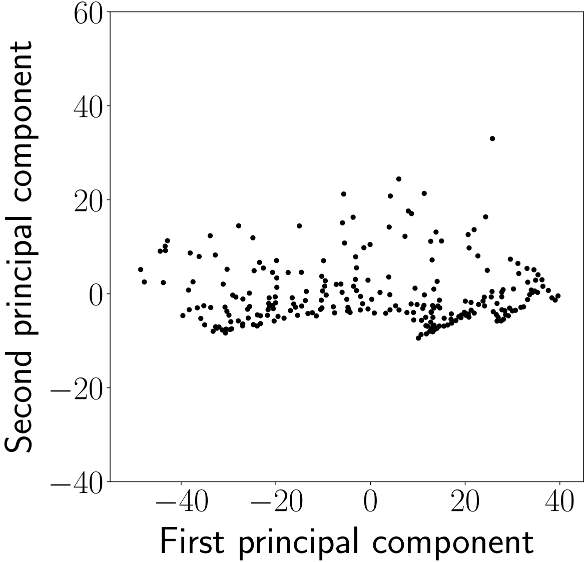
Second principal component



Data



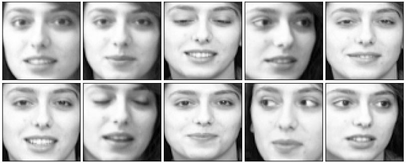
Principal components



Faces

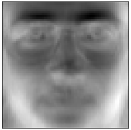
64×64 images from 40 subjects

Vectorized images interpreted as 4096-dimensional vectors



Sample mean

Principal directions

u_1	u_2	u_3	u_4	u_5
				
18.8	11.1	6.30	3.95	2.86

Principal directions

u_{10}	u_{20}	u_{30}	u_{40}	u_{50}
1.32	0.591	0.349	0.217	0.162

Principal directions

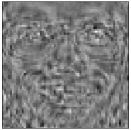
u_{100}

0.061

u_{200}

0.019

u_{250}

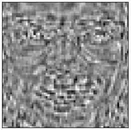


0.011

u_{300}

0.008

u_{350}



0.004

What have we learned

How to find the directions of maximum variance of a random vector or a dataset