The Covariance Matrix

Probability and Statistics for Data Science

Carlos Fernandez-Granda

NYU COURANT INSTITUTE OF MATHEMATICAL SCIENCES

NYU DATA SCIENCE

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

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

Mean of a random vector

The d-dimensional mean of a random vector $\tilde{x}$ is

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

Gaussian random vector

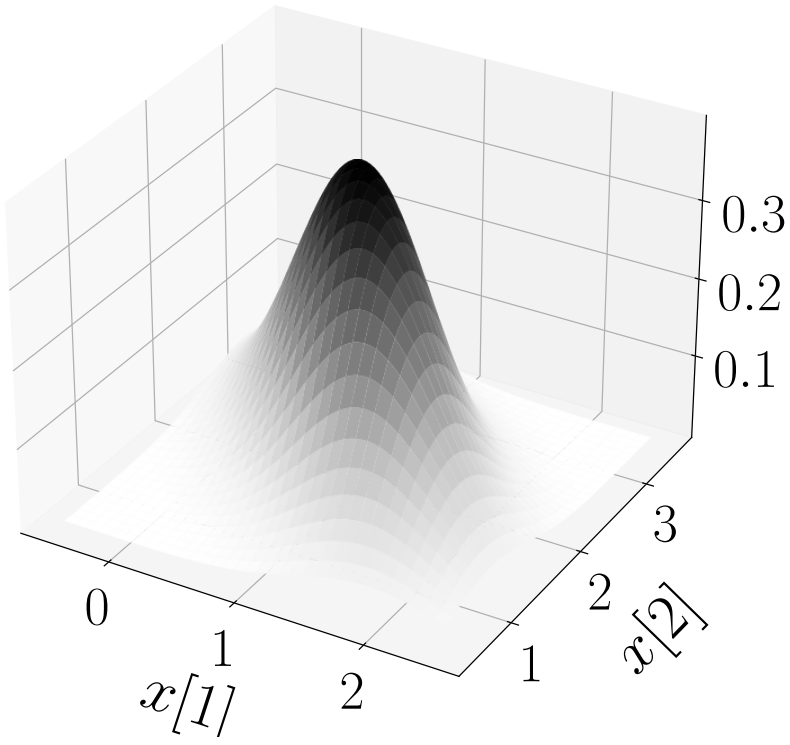
A d-dimensional Gaussian random vector $\tilde{x}$ is a random vector with joint pdf

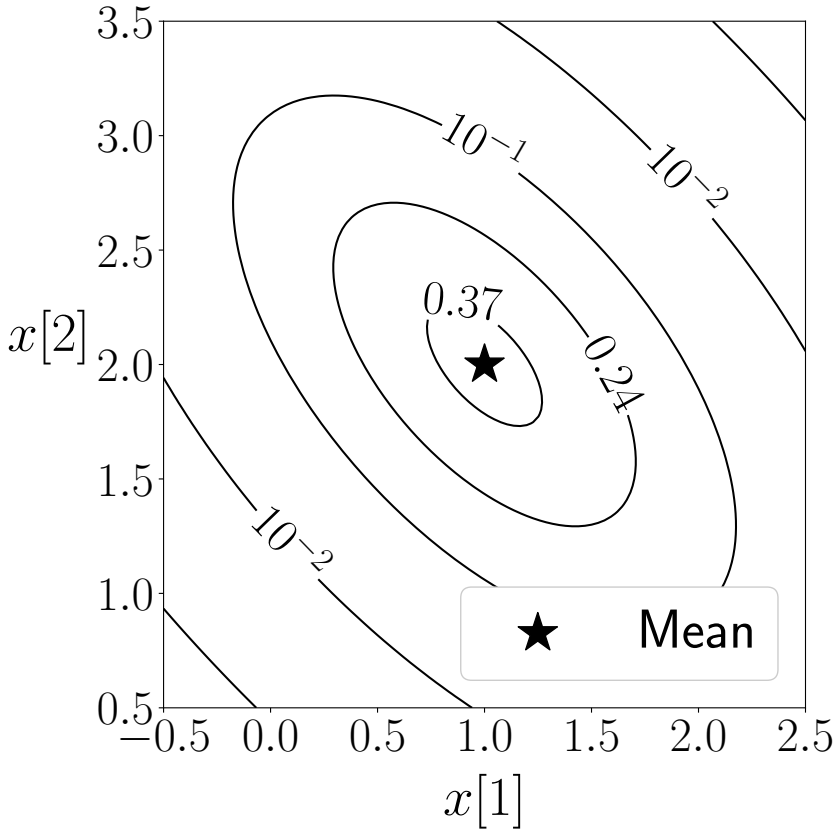
$$f_{\tilde{x}}(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

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

$$\mathrm{E}[\tilde{x}] = \mu$$

Gaussian random vector



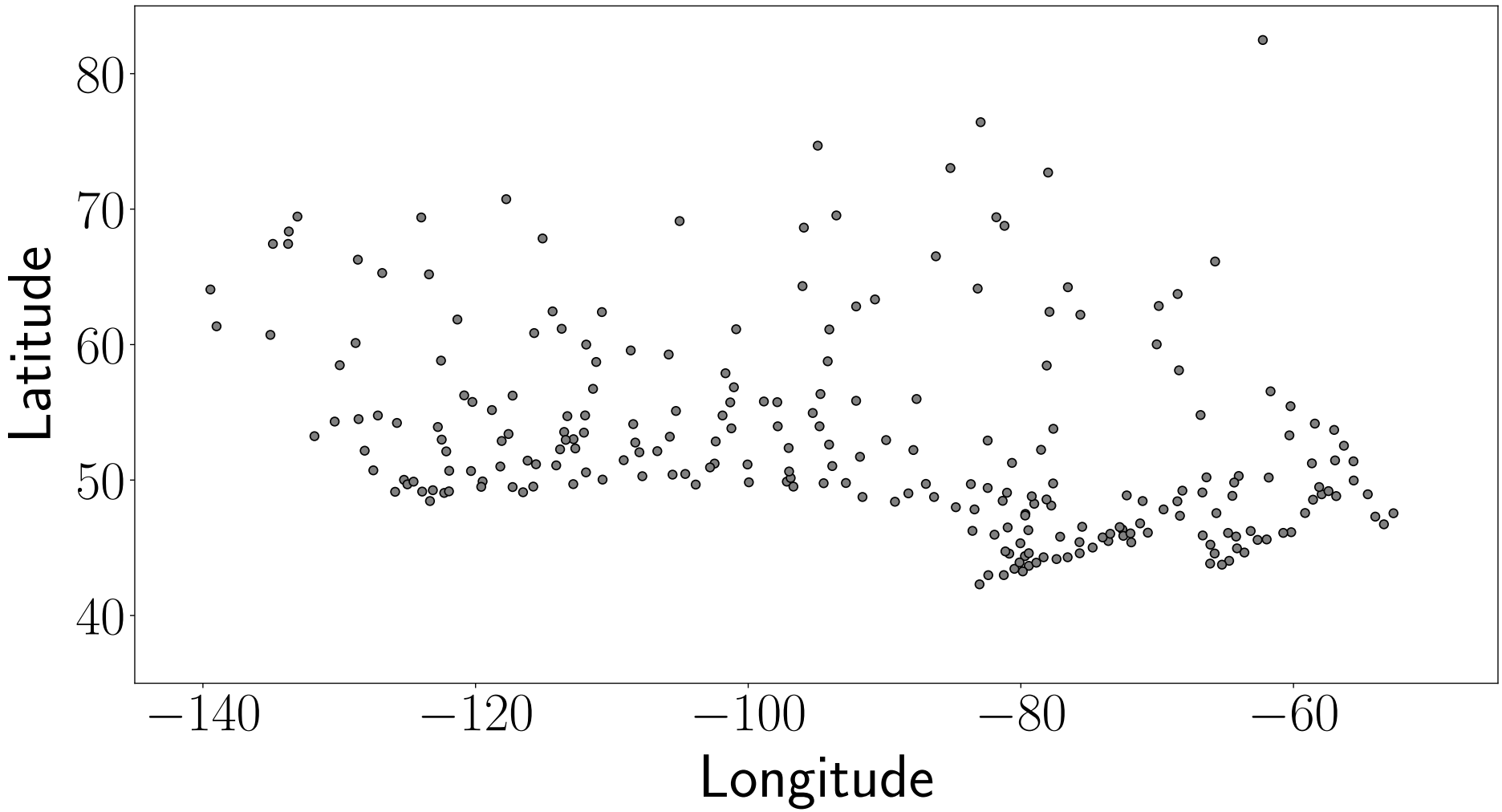


Sample mean

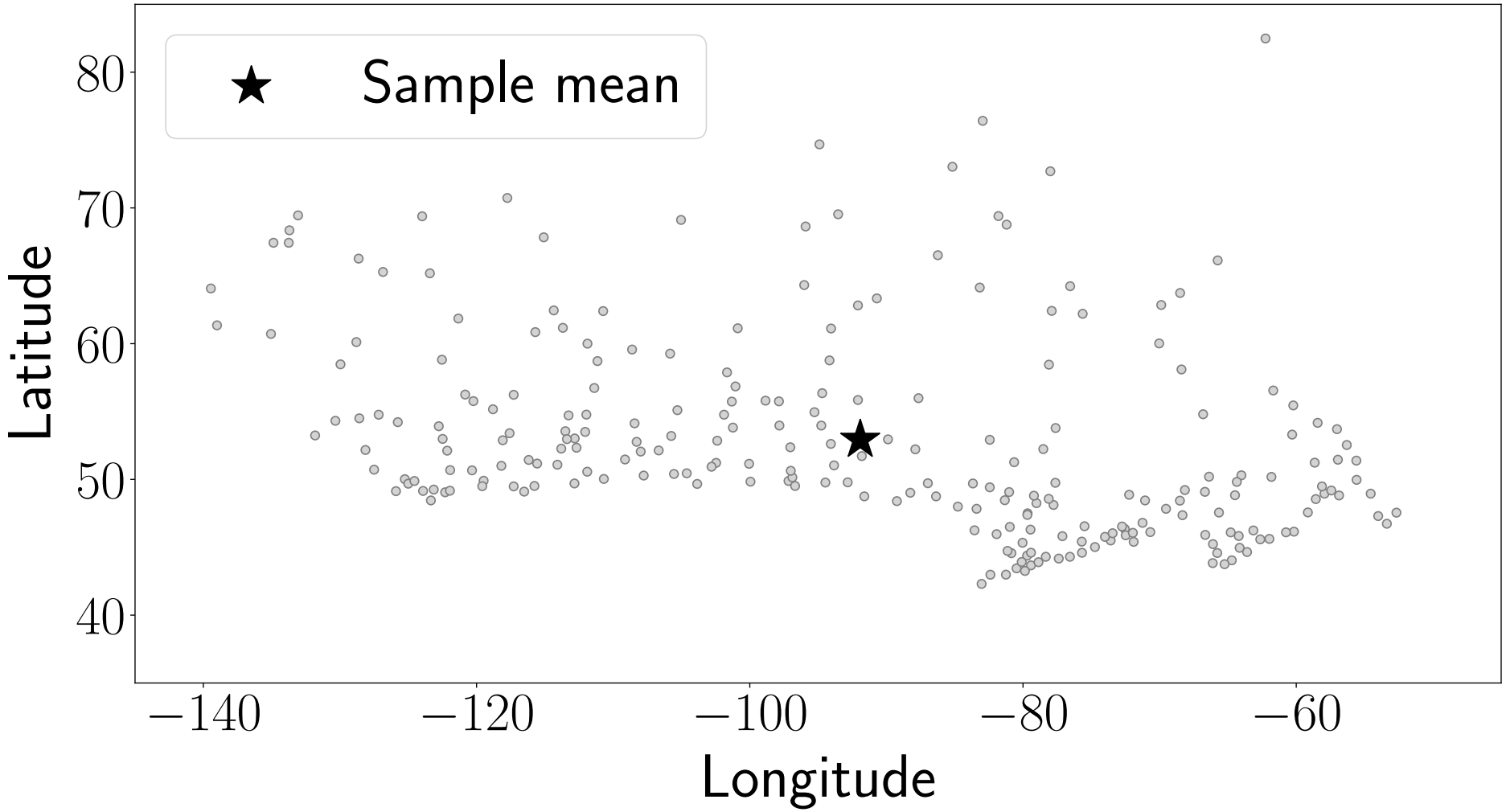
Dataset with d features: $X := \{x_1, x_2, \ldots, x_n\}$

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

Canadian cities



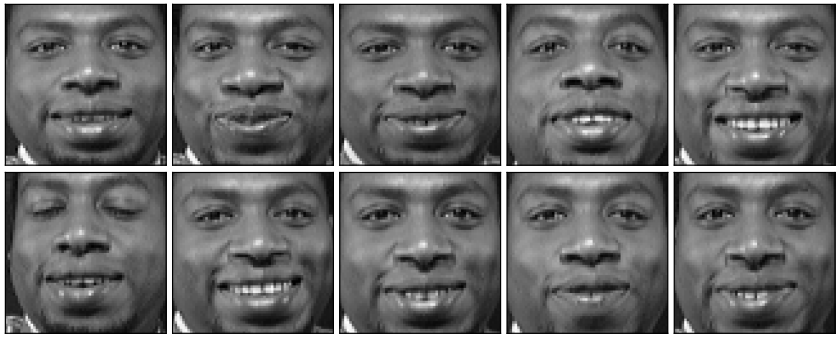
Canadian cities

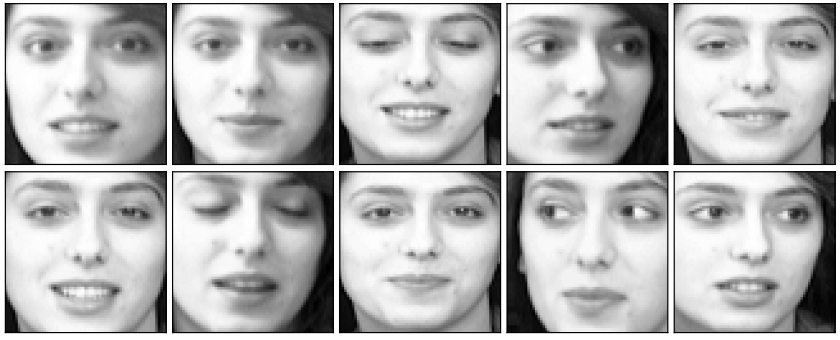


Faces

64×64 images from 40 subjects

Vectorized images interpreted as vectors in $\mathbb{R}^{4096}$





Sample mean

Mean of a random matrix

The mean of a $d_1 \times d_2$ matrix with random entries $\widetilde{M}$ is

$$\mathrm{E}[\widetilde{M}] := \begin{bmatrix} \mathrm{E}\left[\widetilde{M}[1,1]\right] & \mathrm{E}\left[\widetilde{M}[1,2]\right] & \cdots & \mathrm{E}\left[\widetilde{M}[1,d_2]\right] \\ \mathrm{E}\left[\widetilde{M}[2,1]\right] & \mathrm{E}\left[\widetilde{M}[2,2]\right] & \cdots & \mathrm{E}\left[\widetilde{M}[2,d_2]\right] \\ & & \cdots & \\ \mathrm{E}\left[\widetilde{M}[d_1,1]\right] & \mathrm{E}\left[\widetilde{M}[d_1,2]\right] & \cdots & \mathrm{E}\left[\widetilde{M}[d_1,d_2]\right] \end{bmatrix}$$

Linearity of expectation

For any random vector $\tilde{x}$ and deterministic matrix A and vector b

$$\mathrm{E}\left[A\tilde{x} + b\right] = A\mathrm{E}\left[\tilde{x}\right] + b$$

For any random matrix $\widetilde{M}$ and deterministic matrices A and B

$$\mathrm{E}[A\widetilde{M} + B] = A\mathrm{E}[\widetilde{M}] + B$$

$$\begin{aligned}
\mathrm{E}\left[A\tilde{x} + b\right][i] &= \mathrm{E}\left[(A\tilde{x} + b)[i]\right] \\
&= \mathrm{E}\left[\sum_{j=1}^{d} A[i,j]\tilde{x}[j] + b[i]\right] \\
&= \sum_{j=1}^{d} A[i,j]\mathrm{E}\left[\tilde{x}[j]\right] + b[i] \\
&= (A\mathrm{E}\left[\tilde{x}\right] + b)[i]
\end{aligned}$$

Variance

The variance characterizes average variation of a random variable

How can we characterize fluctuations of a random vector?

Variance of linear combinations of the entries

Variance of linear combination $a^T \tilde{x}$?

For any deterministic vector a

$$\begin{aligned}
\operatorname{Var}\left[a^T \tilde{x}\right] &= \mathrm{E}\left[\left(a^T \tilde{x} - \mathrm{E}\left[a^T \tilde{x}\right]\right)^2\right] \\
&= \mathrm{E}\left[\left(a^T \left(\tilde{x} - \mathrm{E}\left[\tilde{x}\right]\right)\right)^2\right] \\
&= \mathrm{E}\left[\left(a^T \operatorname{ct}\left(\tilde{x}\right)\right)^2\right] \\
&= \mathrm{E}\left[a^T \operatorname{ct}\left(\tilde{x}\right) \operatorname{ct}\left(\tilde{x}\right)^T a\right] \\
&= a^T \mathrm{E}\left[\operatorname{ct}\left(\tilde{x}\right) \operatorname{ct}\left(\tilde{x}\right)^T\right] a
\end{aligned}$$

where $\operatorname{ct}\left(\tilde{x}\right) := \tilde{x} - \mathrm{E}\left[\tilde{x}\right]$

$$\mathrm{E}\left[\mathrm{ct}\left(\tilde{x}\right)\mathrm{ct}\left(\tilde{x}\right)^{T}\right]$$

Diagonal entries

$$\begin{aligned}\mathrm{E}\left[\left(\mathrm{ct}\left(\tilde{x}\right)\mathrm{ct}\left(\tilde{x}\right)^{T}\right)[i,i]\right] &= \mathrm{E}\left[\mathrm{ct}\left(\tilde{x}[i]\right)^{2}\right] \\ &= \mathrm{Var}\left[\tilde{x}[i]\right]\end{aligned}$$

Off-diagonal entries

$$\begin{aligned}\mathrm{E}\left[\left(\mathrm{ct}\left(\tilde{x}\right)\mathrm{ct}\left(\tilde{x}\right)^{T}\right)[i,j]\right] &= \mathrm{E}\left[\mathrm{ct}\left(\tilde{x}[i]\right)\mathrm{ct}\left(\tilde{x}[j]\right)\right] \\ &= \mathrm{Cov}\left[\tilde{x}[i],\tilde{x}[j]\right]\end{aligned}$$

Covariance matrix

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

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

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

Variance of linear combination $a^T \tilde{x}$

For any deterministic vector a

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

Cheese sandwich

Ingredients: Bread, local cheese, and imported cheese

Prices: Random vector $\tilde{x}$ (cents/gram) with covariance matrix

$$\Sigma_{\tilde{x}} = \begin{bmatrix} 1 & 0.8 & 0 \\ 0.8 & 1 & 0 \\ 0 & 0 & 1.2 \end{bmatrix}$$

Two recipes:

1. 100g bread, 50g local cheese, and 50g imported cheese

2. 100g bread, 100g local cheese, and no imported cheese

Which has higher standard deviation?

Recipe 1

$$\sigma_{100\tilde{x}[1]+50\tilde{x}[2]+50\tilde{x}[3]} = \sqrt{\begin{bmatrix}100 & 50 & 50\end{bmatrix} \Sigma_{\tilde{x}} \begin{bmatrix}100\\50\\50\end{bmatrix}}$$

$$= \sqrt{\begin{bmatrix}100 & 50 & 50\end{bmatrix} \begin{bmatrix}1 & 0.8 & 0\\0.8 & 1 & 0\\0 & 0 & 1.2\end{bmatrix} \begin{bmatrix}100\\50\\50\end{bmatrix}}$$

$$= 153 \text{ cents}$$

Recipe 2

$$\sigma_{100\tilde{x}[1]+100\tilde{x}[2]} = \sqrt{\begin{bmatrix}100 & 100 & 0\end{bmatrix} \Sigma_{\tilde{x}} \begin{bmatrix}100\\100\\0\end{bmatrix}}$$

$$= \sqrt{\begin{bmatrix}100 & 100 & 0\end{bmatrix} \begin{bmatrix}1 & 0.8 & 0\\0.8 & 1 & 0\\0 & 0 & 1.2\end{bmatrix} \begin{bmatrix}100\\100\\0\end{bmatrix}}$$

$$= 190 \text{ cents}$$

Gaussian random vector

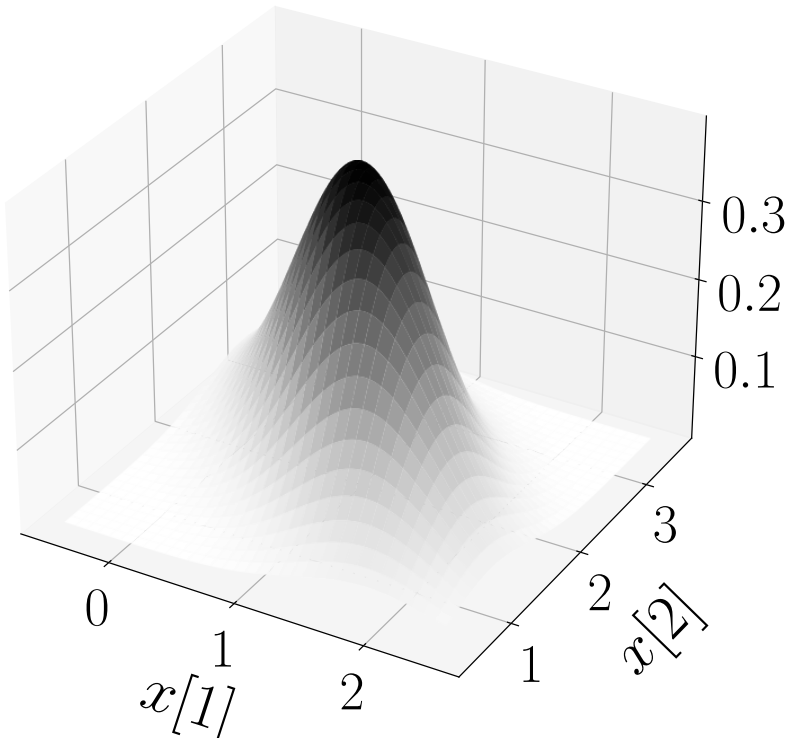
A d-dimensional Gaussian random vector $\tilde{x}$ is a random vector with joint pdf

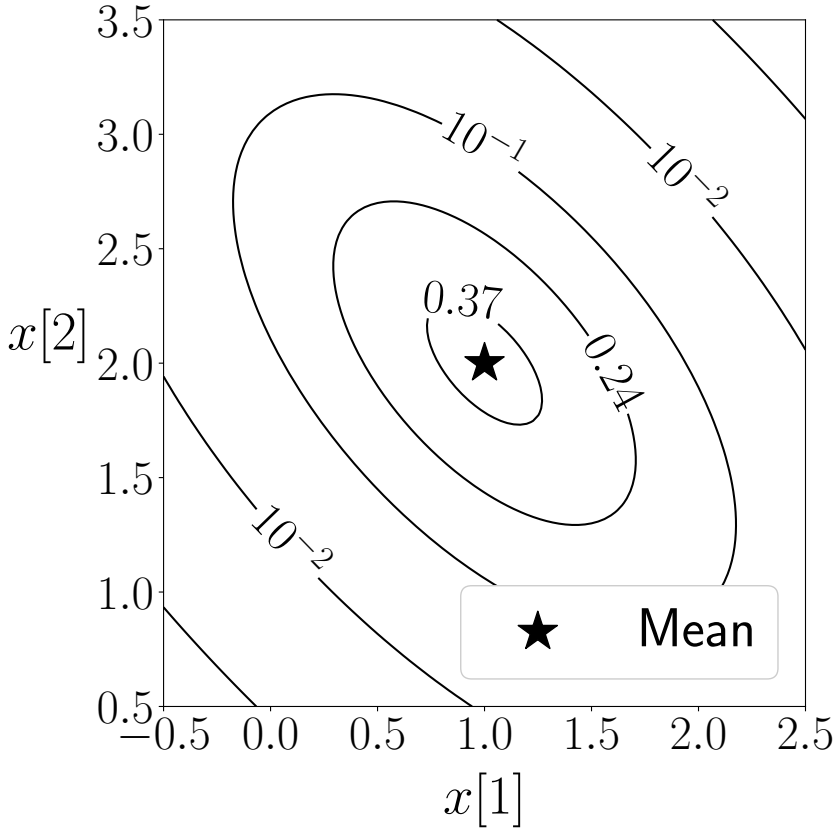
$$f_{\tilde{x}}(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

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

$$\Sigma_{\tilde{x}} = \Sigma$$

Gaussian random vector





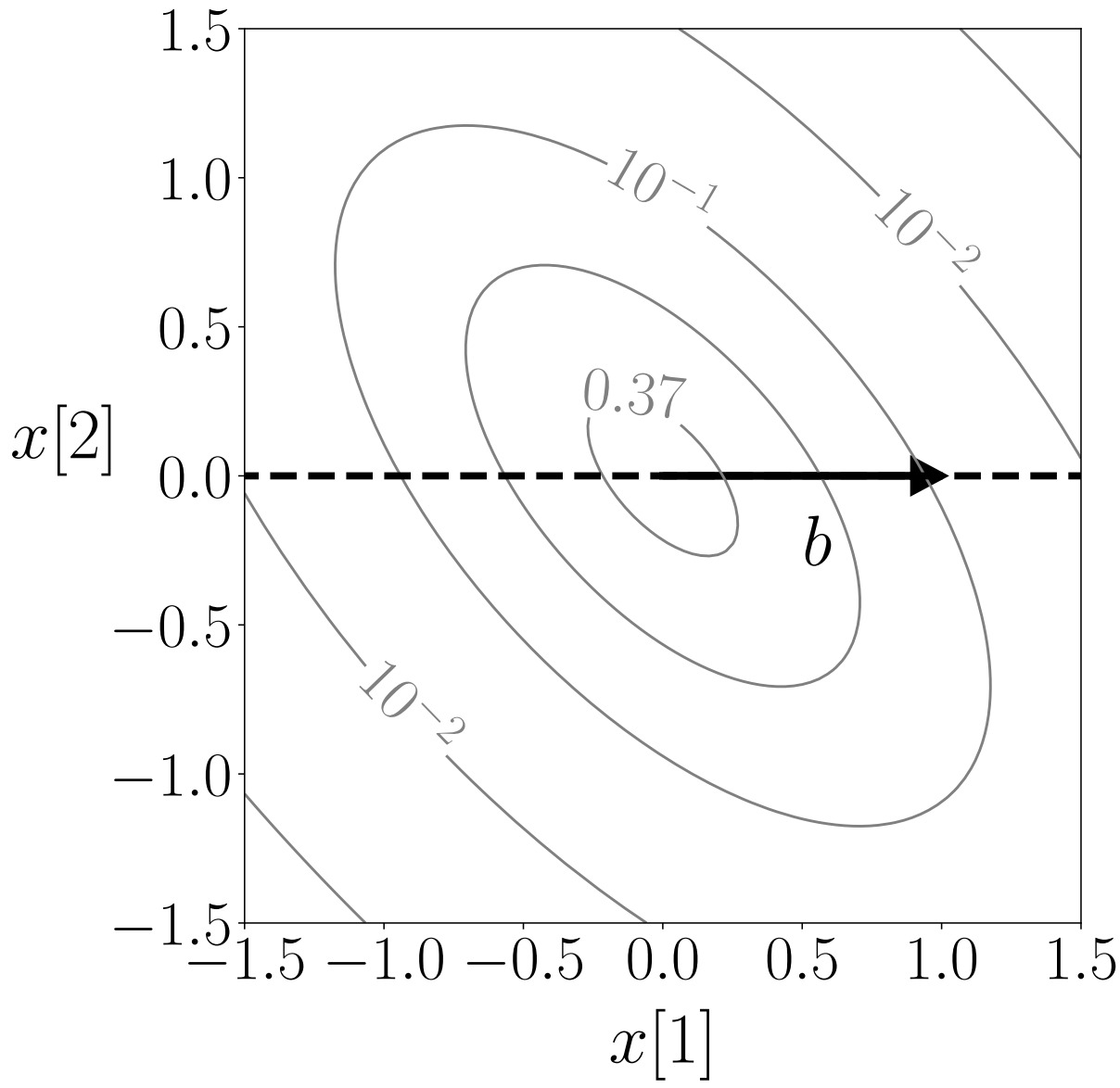
$$\Sigma_{\tilde{x}} := \begin{bmatrix} 0.5 & -0.3 \\ -0.3 & 0.5 \end{bmatrix}$$

Variance in a certain direction?

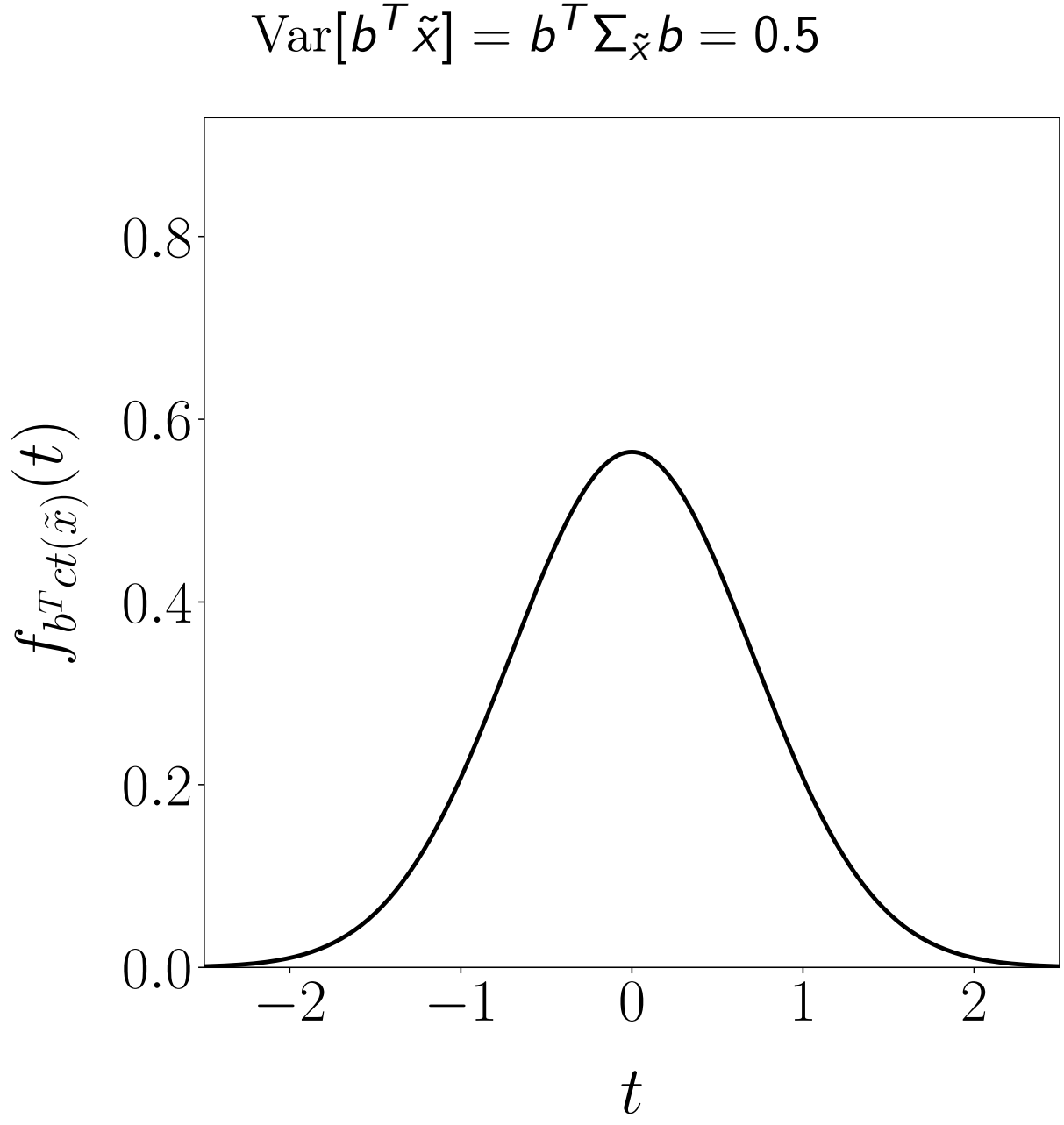
After centering by subtracting the mean

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

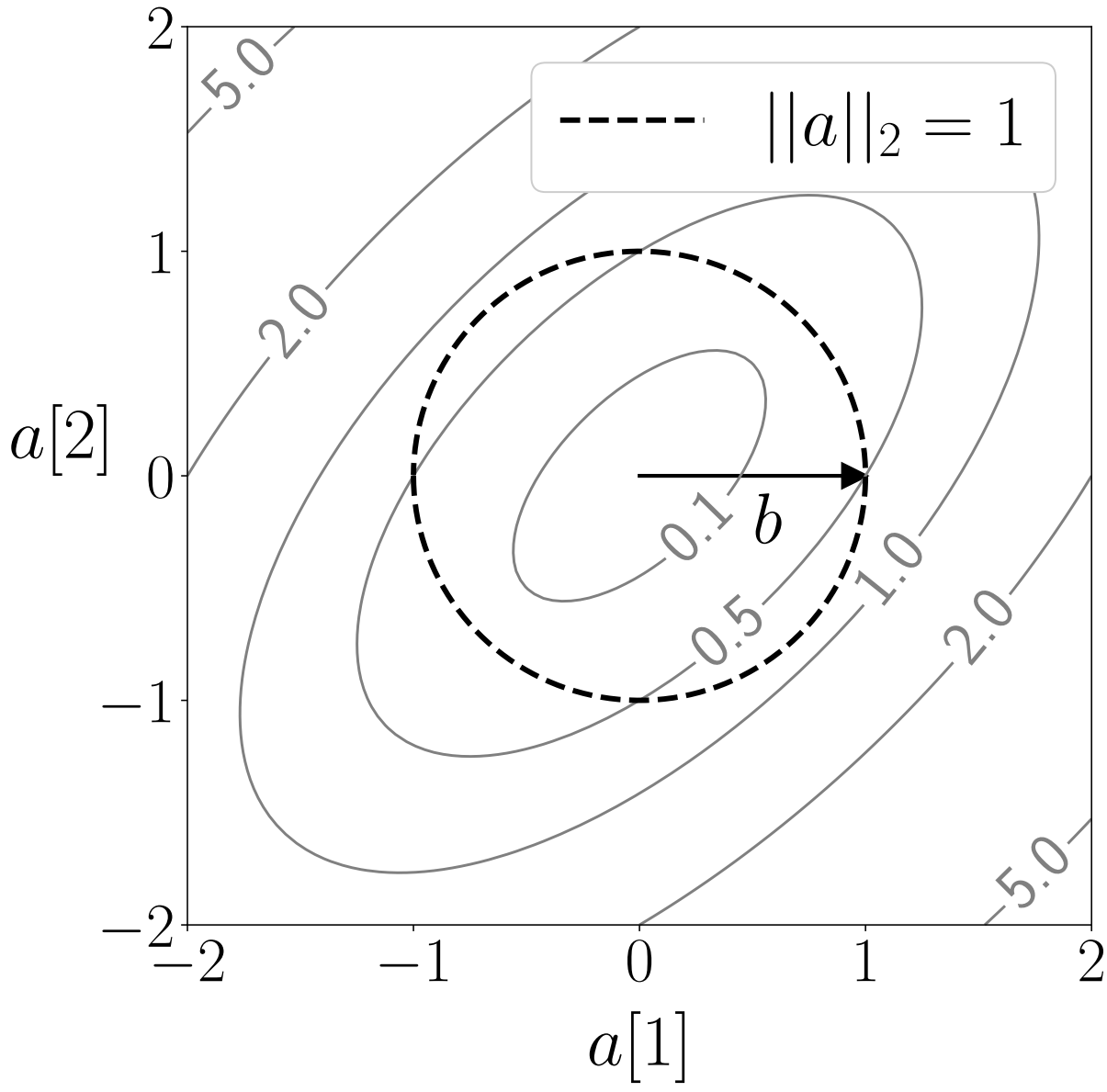
Variance in a certain direction?



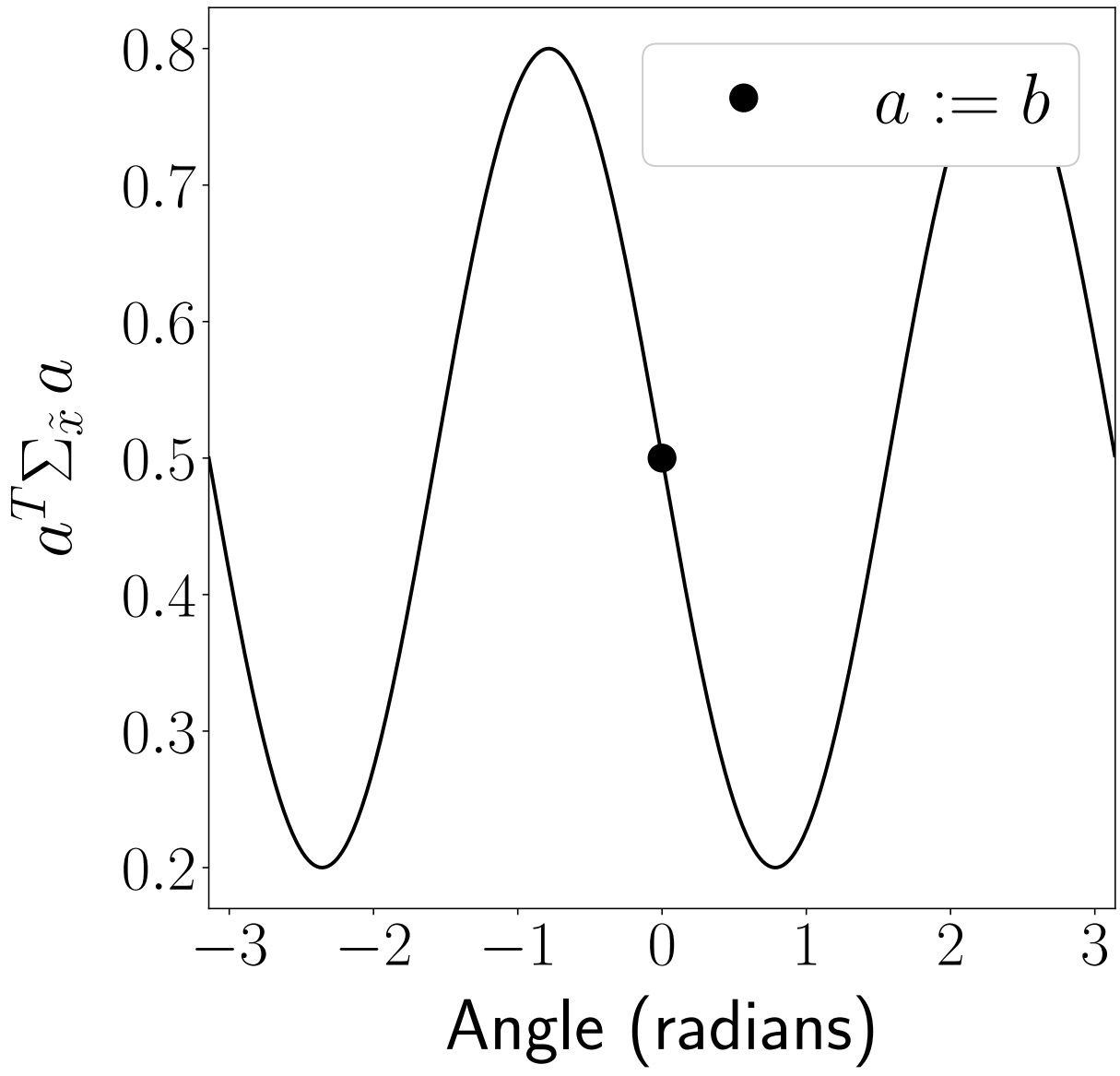
Pdf of $b^T\tilde{x}$



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



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



Covariance matrix of a dataset

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

jth feature: $X[j] := \{x_1[j], \ldots, x_n[j]\}$

$v(X[j])$: sample variance of $X[j]$

$c(X[j], X[k])$: sample covariance of $X[j]$ and $X[k]$

Covariance matrix of a dataset

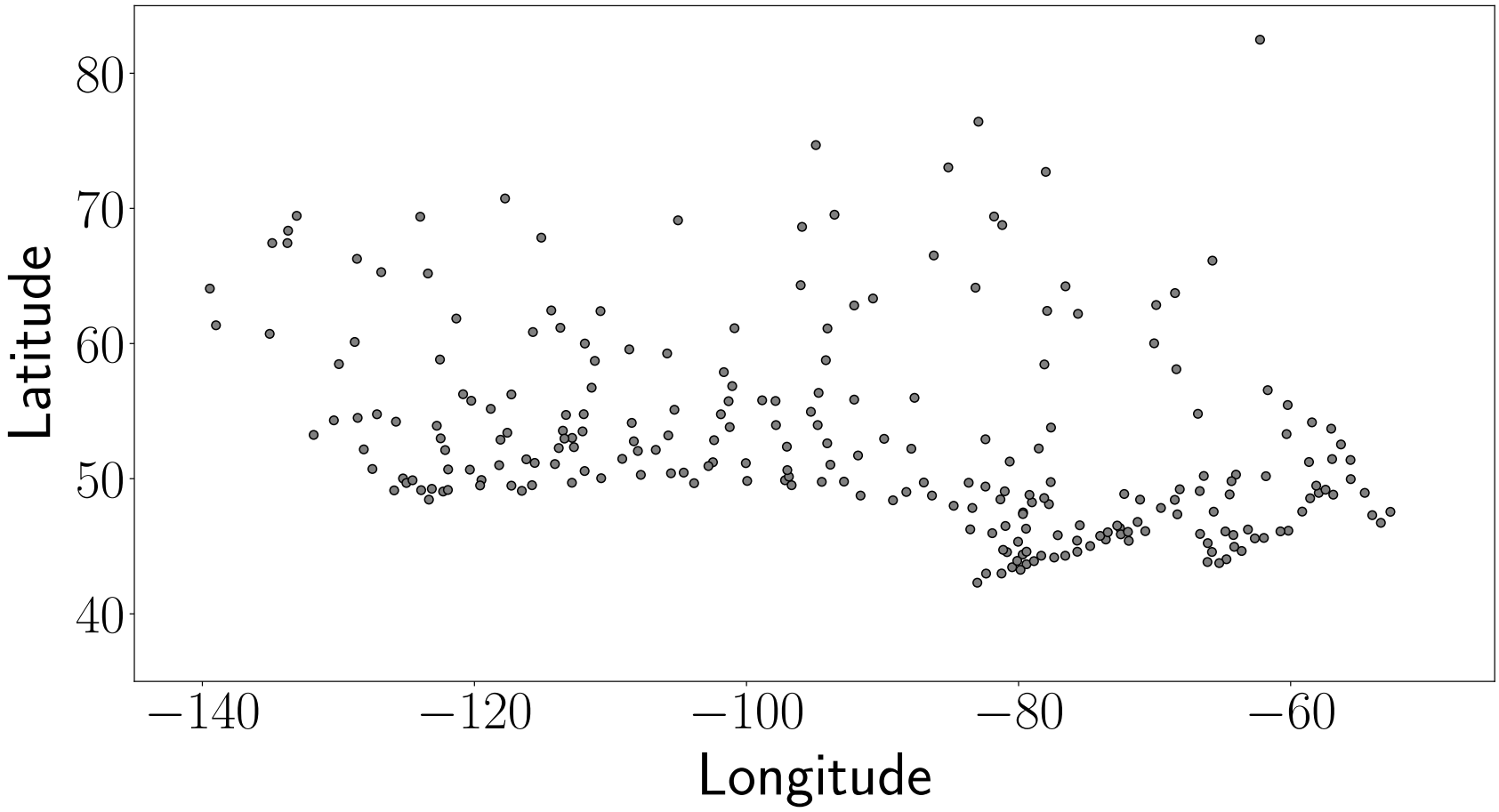
Data with d features: $X := \{x_1, x_2, \ldots, x_n\}$

Sample covariance matrix of X:

$$\Sigma_X := \begin{bmatrix} v(X[1]) & c(X[1], X[2]) & \cdots & c(X[1], X[d]) \\ c(X[1], X[2]) & v(X[2]) & \cdots & c(X[2], X[d]) \\ \vdots & \vdots & \ddots & \vdots \\ c(X[1], X[d]) & c(X[2], X[d]) & \cdots & v(X[d]) \end{bmatrix}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} \mathrm{ct}\,(x_i)\,\mathrm{ct}\,(x_i)^T \qquad \mathrm{ct}\,(x_i) := x_i - m(X)$$

Cities in Canada



Sample covariance matrix:

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

Sample covariance matrix of temperature data

	Tucson, AZ	Hilo, HI	Durham, NC	Ithaca, NY
Tucson, AZ	78.6	14.7	54.8	65.0
Hilo, HI	14.7	8.4	9.5	11.8
Durham, NC	54.8	9.5	89.4	97.4
Ithaca, NY	65.0	11.8	97.4	137.3

Sample correlation matrix

	Tucson, AZ	Hilo, HI	Durham, NC	Ithaca, NY
Tucson, AZ	1	0.57	0.65	0.63
Hilo, HI	0.57	1	0.35	0.35
Durham, NC	0.65	0.35	1	0.88
Ithaca, NY	0.63	0.35	0.88	1

Sample variance of linear combination

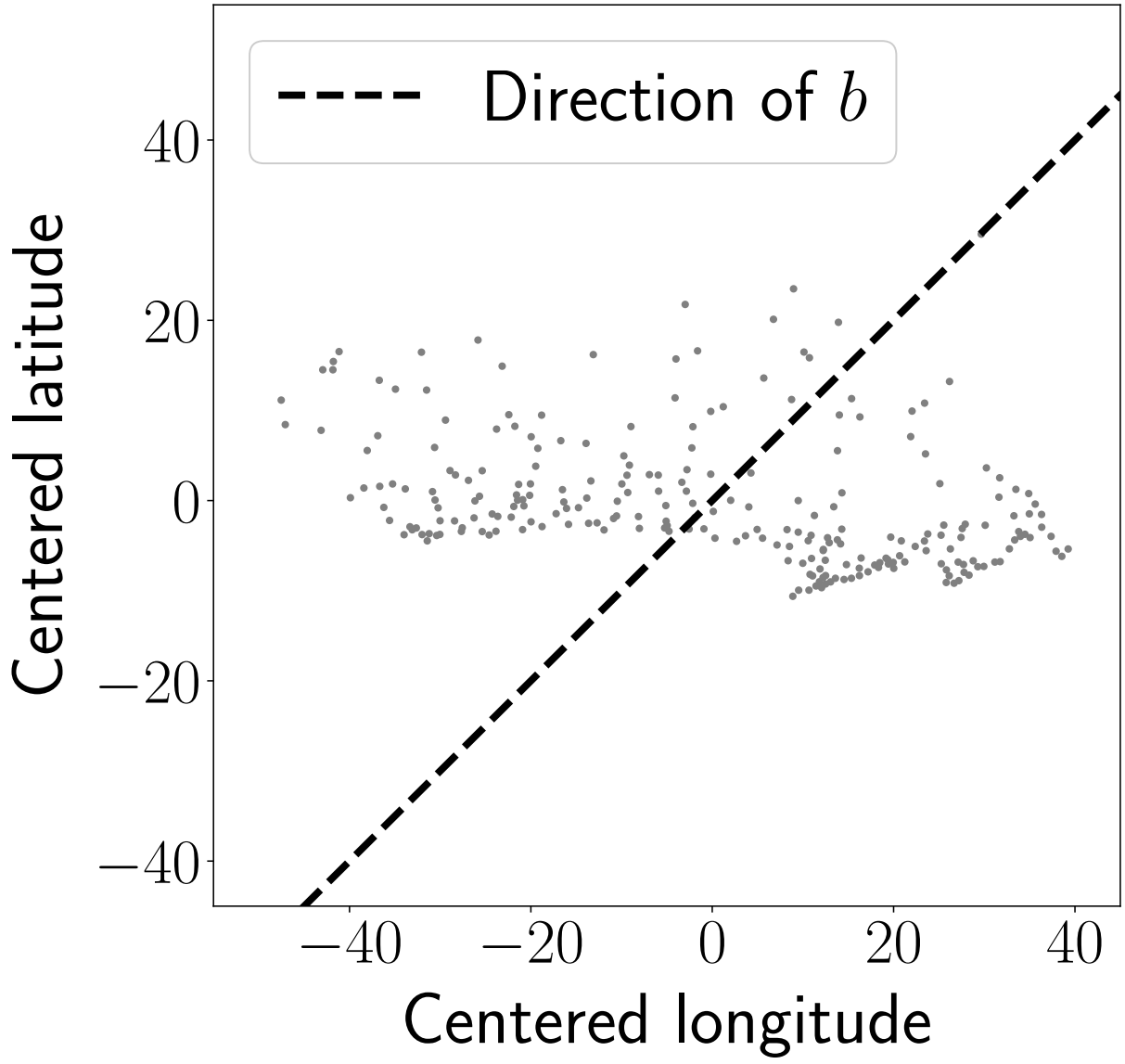
Dataset: $X = \{x_1, \ldots, x_n\}$

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

Sample variance of linear combination

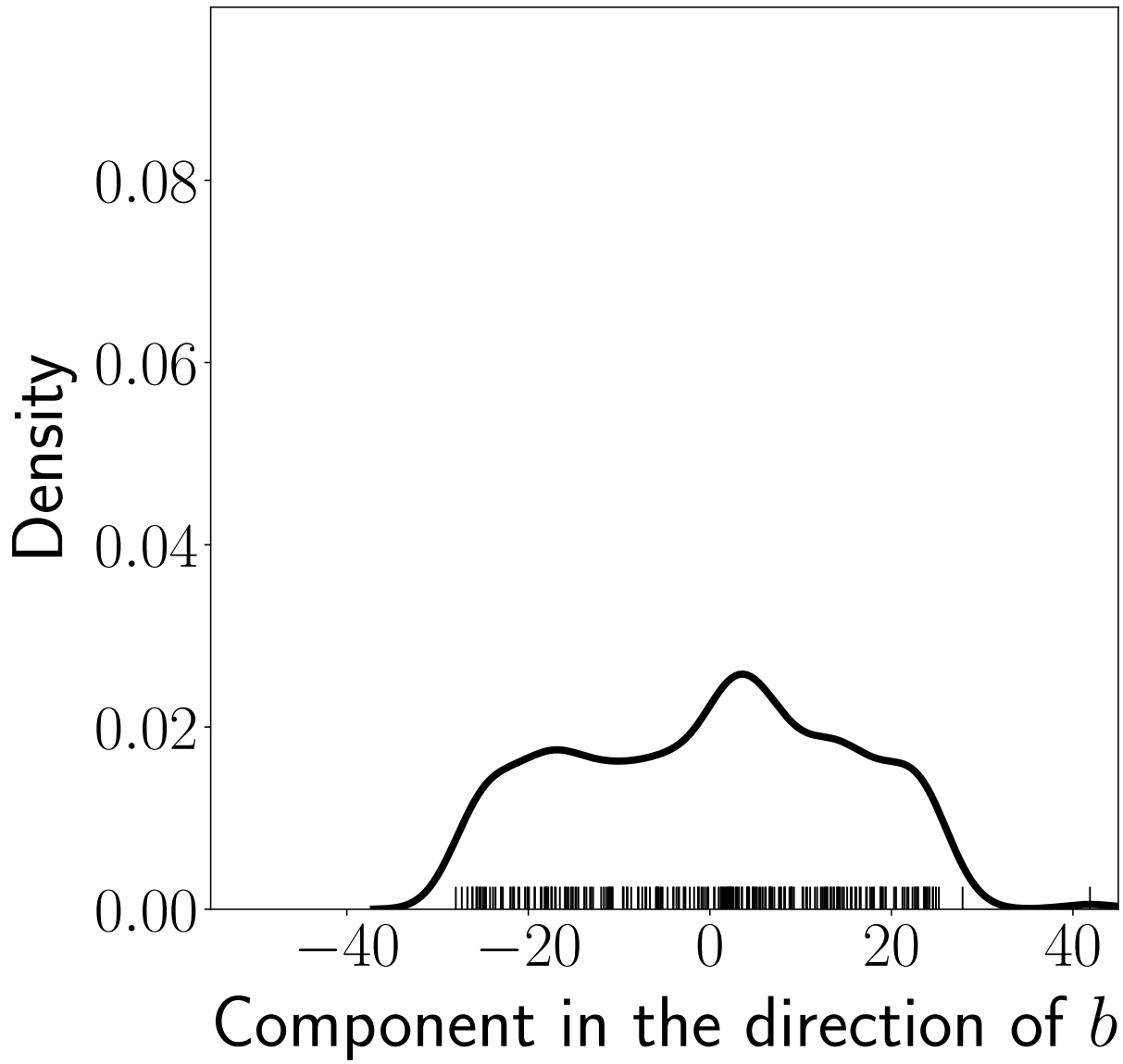
$$
\begin{aligned}
v\left(X_a\right) &= \frac{1}{n-1} \sum_{i=1}^{n} \left(a^T x_i - \frac{1}{n} \sum_{j=1}^{n} a^T x_j \right)^2 \\
&= \frac{1}{n-1} \sum_{i=1}^{n} \left(a^T \left(x_i - \frac{1}{n} \sum_{j=1}^{n} x_j \right) \right)^2 \\
&= \frac{1}{n-1} \sum_{i=1}^{n} (a^T \operatorname{ct}\left(x_i\right))^2 \\
&= \frac{1}{n-1} \sum_{i=1}^{n} a^T \operatorname{ct}\left(x_i\right) \operatorname{ct}\left(x_i\right)^T a \\
&= a^T \left(\frac{1}{n-1} \sum_{i=1}^{n} \operatorname{ct}\left(x_i\right) \operatorname{ct}\left(x_i\right)^T \right) a \\
&= a^T \Sigma_X a
\end{aligned}
$$

Variance in a certain direction?

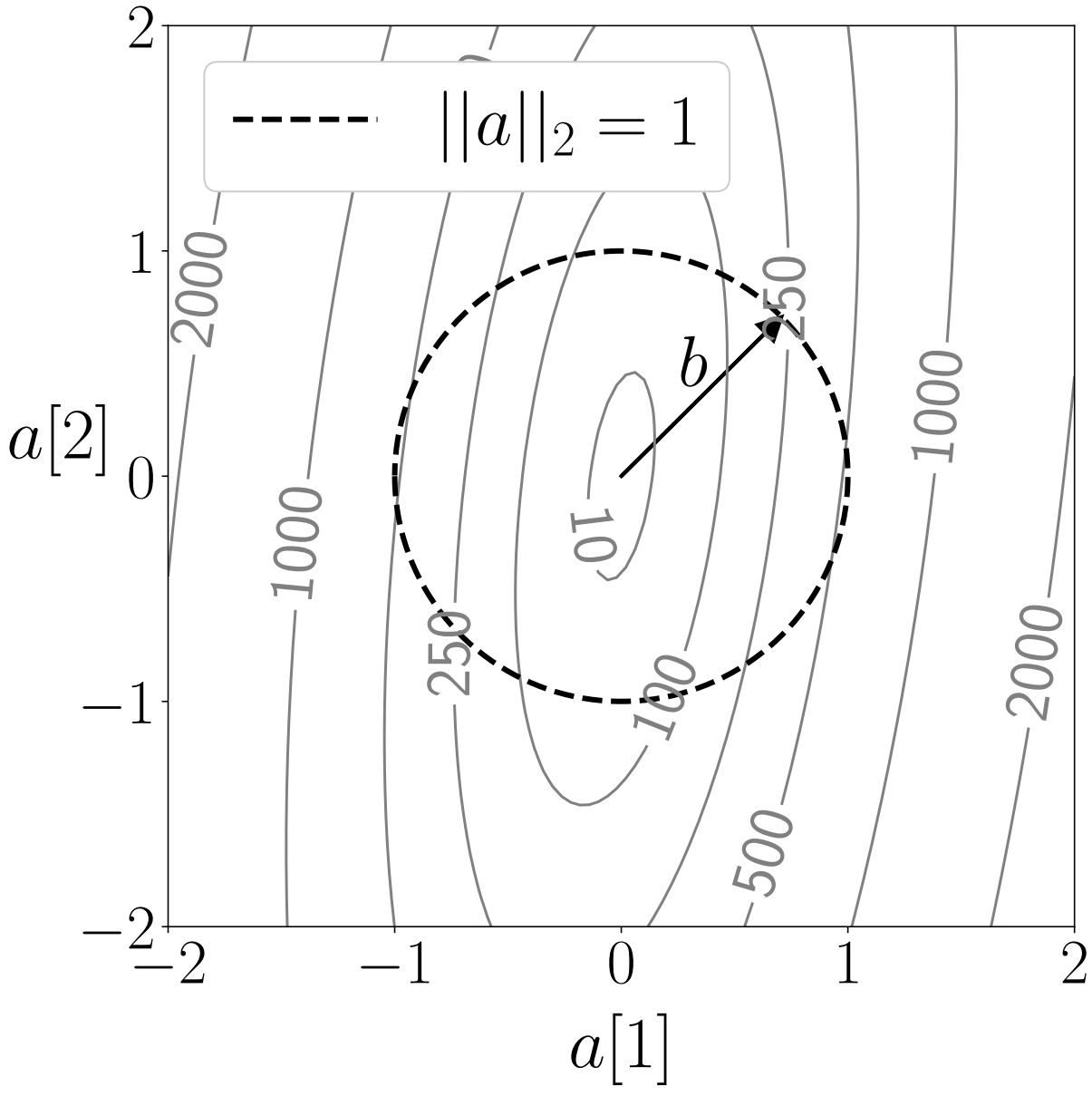


Estimated pdf in the direction of b

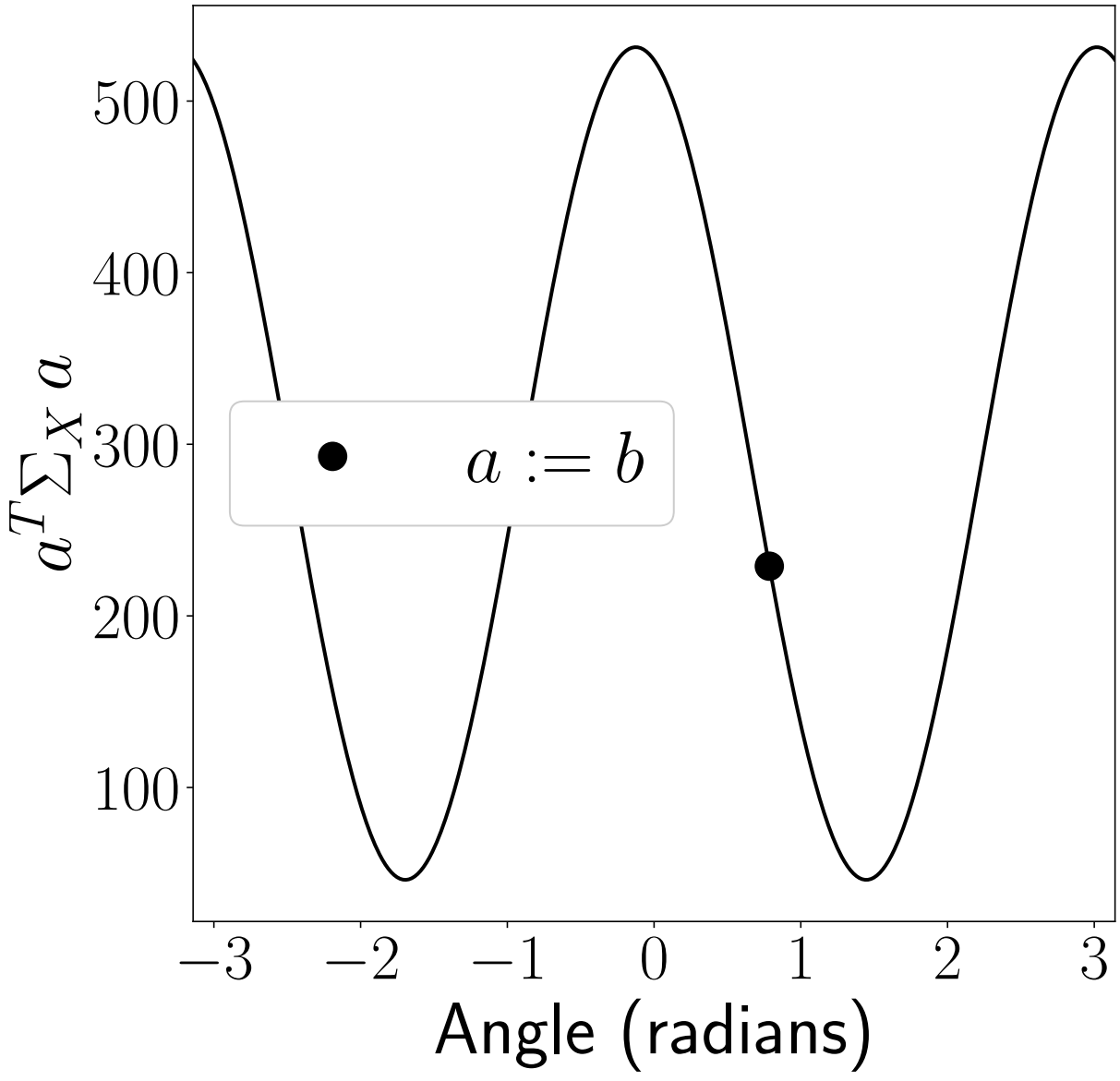
$$v(X_b) = b^T \Sigma_X b = 229$$



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



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



What have we learned

Mean of vectors and matrices

Covariance matrix encodes variance of linear combinations

How to estimate covariance matrix from data

Sample covariance matrix encodes sample variance of linear combinations