Effect of Data Rescaling

- Rescaling of the data affects the results of LPP and PCA, because Euclidean distance depends on the scale of data.

$$\boldsymbol{B}_{PCA} = \underset{\boldsymbol{B} \in \mathbb{R}^{m \times d}}{\operatorname{argmin}} \left[\sum_{i=1}^{n} \| \boldsymbol{B}^\top \boldsymbol{B} \boldsymbol{x}_i - \boldsymbol{x}_i \|^2 \right]$$

$$\text{subject to } \boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{I}_m$$

$$\boldsymbol{B}_{LPP} = \underset{\boldsymbol{B} \in \mathbb{R}^{m \times d}}{\operatorname{argmin}} \left[\sum_{i,j=1}^{n} \| \boldsymbol{B}\boldsymbol{x}_i - \boldsymbol{B}\boldsymbol{x}_j \|^2 \boldsymbol{W}_{i,j} \right],$$

$$\text{subject to } \boldsymbol{B}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^\top \boldsymbol{B}^\top = \boldsymbol{I}_m$$

Invariance under Data Rescaling

- To be invariant under data rescaling, embedding criterion should not depend on the scale of the data.
- Idea: Use data distributions, rather than distances.
- Suppose data samples are i.i.d. random variables.

$$\boldsymbol{x}_i \overset{i.i.d.}{\sim} p(\boldsymbol{x})$$

Gaussian Distribution

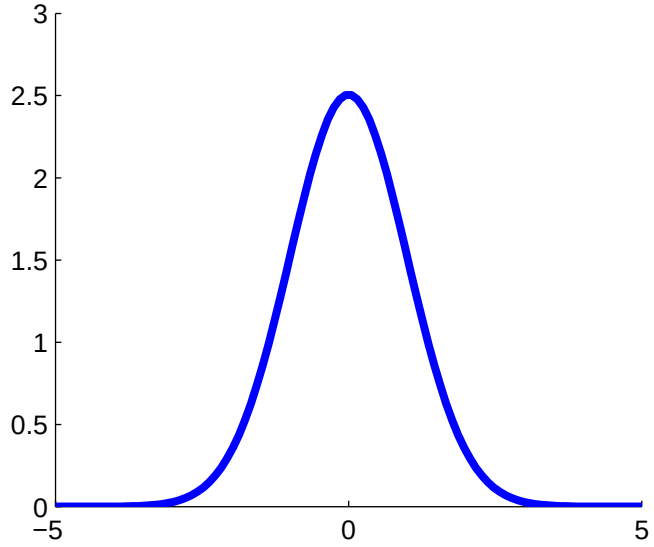
- Gaussian distribution: Probability density function is given by

$$\phi_{\boldsymbol{\theta},\boldsymbol{\Gamma}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Gamma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\theta})^{\top}\boldsymbol{\Gamma}^{-1}(\boldsymbol{x}-\boldsymbol{\theta})\right)$$

- $\boldsymbol{\theta}, \boldsymbol{\Gamma}$:Mean, covariance

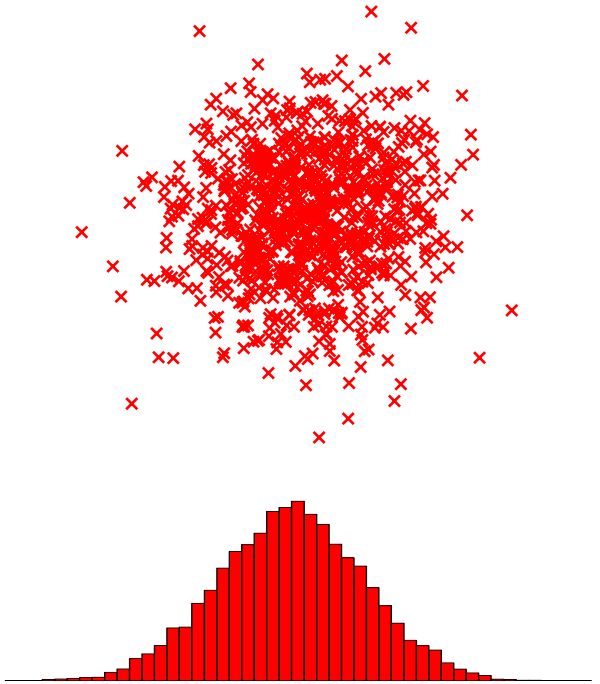
$$\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\theta}$$

$$\mathbb{E}[(\boldsymbol{x}-\theta)(\boldsymbol{x}-\theta)^{\top}] = \boldsymbol{\Gamma}$$



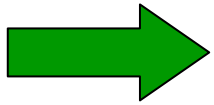
Interesting Directions for Data Visualization

- What distribution is interesting to visualize?
- If data follows the Gaussian distribution, samples are spherically distributed.
- Visualizing spherically distributed samples is not so interesting.
- What about "non-Gaussian" data?



Non-Gaussian Distributed Data

- Non-Gaussian data look interesting:

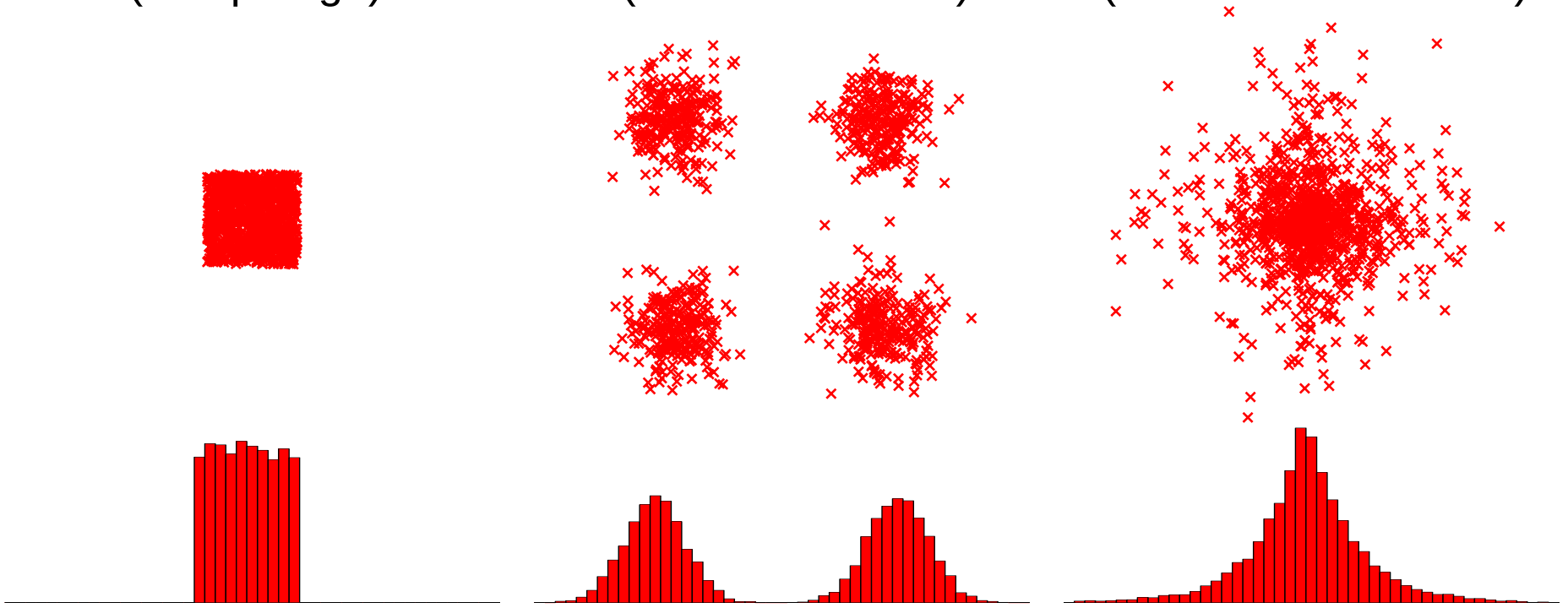


We want to project the data so that it has non-Gaussian distributions

Uniform (sharp edge)

Gaussian mixture (cluster structure)

Laplacian (existence of outliers)



Projection Pursuit

- Idea: Iteratively find non-Gaussian directions in the data
- For $k = 1, 2, \ldots, m$

  Find the most non-Gaussian direction in data:

$$\boldsymbol{\psi}_k = \underset{\boldsymbol{b} \perp \{\boldsymbol{\psi}_i\}_{i=1}^{k-1}}{\operatorname{argmax}} J_{PP}(\boldsymbol{b}) \quad \text{subject to } \|\boldsymbol{b}\| = 1$$

- PP embedding of a sample $\boldsymbol{x}'$:

$$\boldsymbol{z}' = \boldsymbol{B}_{PP}\boldsymbol{x}' \qquad \boldsymbol{B}_{PP} = (\boldsymbol{\psi}_1 | \boldsymbol{\psi}_2 | \cdots | \boldsymbol{\psi}_m)^\top$$

Kurtosis

- PP needs a non-Gaussianity measure.
- Kurtosis:

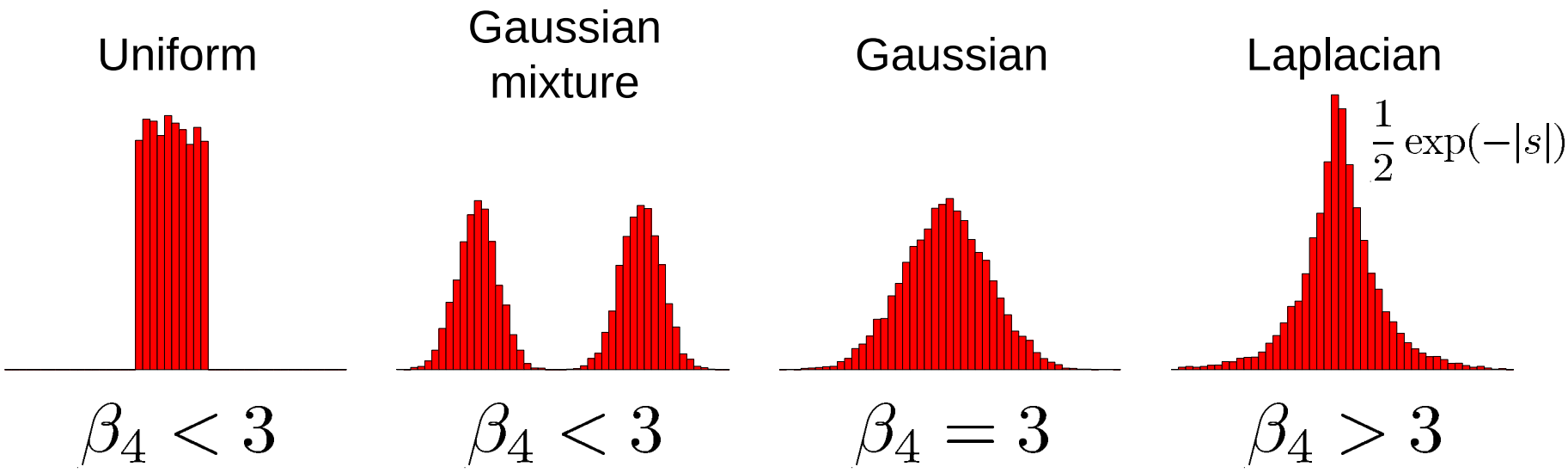
$$\beta_4 = \frac{\mathbb{E}[(s - \mathbb{E}[s])^4]}{(\mathbb{E}[(s - \mathbb{E}[s])^2])^2} \quad (> 0)$$

$$s = \langle \boldsymbol{b}, \boldsymbol{x} \rangle$$

- If tail of distribution is
  - Heavy ⇒ β_4 is large
  - Light ⇒ β_4 is small

Kurtosis (cont.)

- $\beta_4 = 3$: Gaussian distribution
- $\beta_4 < 3$: Sub-Gaussian distribution
- $\beta_4 > 3$: Super-Gaussian distribution



Kurtosis-Based Non-Gaussianity Measure

- Non-Gaussianity is strong if $(\beta_4 - 3)^2$ is large.
- In practice, we use empirical approximation:

$$J_{PP}(\boldsymbol{b}) = \left(\frac{\frac{1}{n}\sum_{i=1}^{n}[(s_i - \overline{s})^4]}{(\frac{1}{n}\sum_{i=1}^{n}[(s_i - \overline{s})^2])^2} - 3 \right)^2 \qquad s_i = \langle \boldsymbol{b}, \boldsymbol{x}_i \rangle \qquad \overline{s} = \frac{1}{n}\sum_{i=1}^{n} s_i$$

- There is no known method for analytically solving the optimization problem.

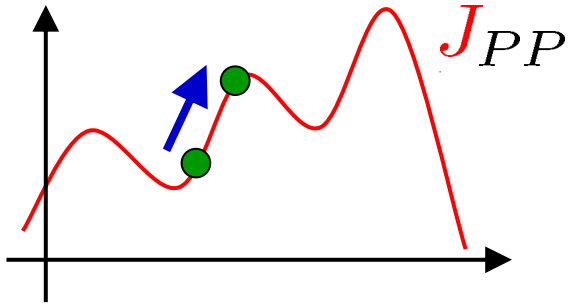
$$\boldsymbol{\psi} = \underset{\boldsymbol{b} \perp \{\boldsymbol{\psi}_i\}_{i=1}^{k-1}}{\operatorname{argmax}} \; J_{PP}(\boldsymbol{b}) \quad \text{subject to } \|\boldsymbol{b}\| = 1$$

- We resort to numerical methods.

Gradient Ascent Approach

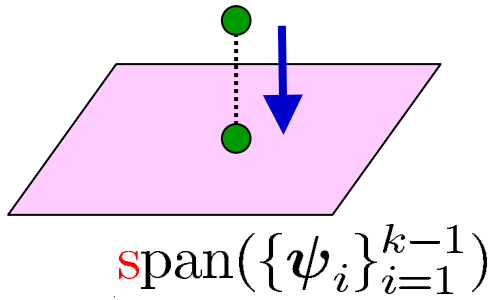
- Repeat the following until convergence:
  - Update $\boldsymbol{b}$ to increase J_{PP} :

$$\boldsymbol{b} \longleftarrow \boldsymbol{b} + \varepsilon \frac{\partial J_{PP}}{\partial \boldsymbol{b}} \qquad (\varepsilon > 0)$$



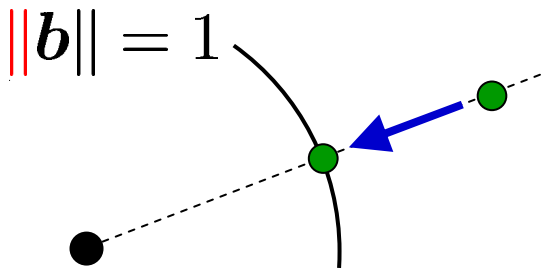
  - Modify $\boldsymbol{b}$ to satisfy $\boldsymbol{b} \perp \{\boldsymbol{\psi}_i\}_{i=1}^{k-1}$:

$$\boldsymbol{b} \longleftarrow \boldsymbol{b} - \sum_{i=1}^{k-1} \langle \boldsymbol{b}, \boldsymbol{\psi}_i \rangle \boldsymbol{\psi}_i$$



  - Modify $\boldsymbol{b}$ to satisfy $\|\boldsymbol{b}\| = 1$:

$$\boldsymbol{b} \longleftarrow \boldsymbol{b} / \|\boldsymbol{b}\|$$



Data Centering and Sphering

- Centering:

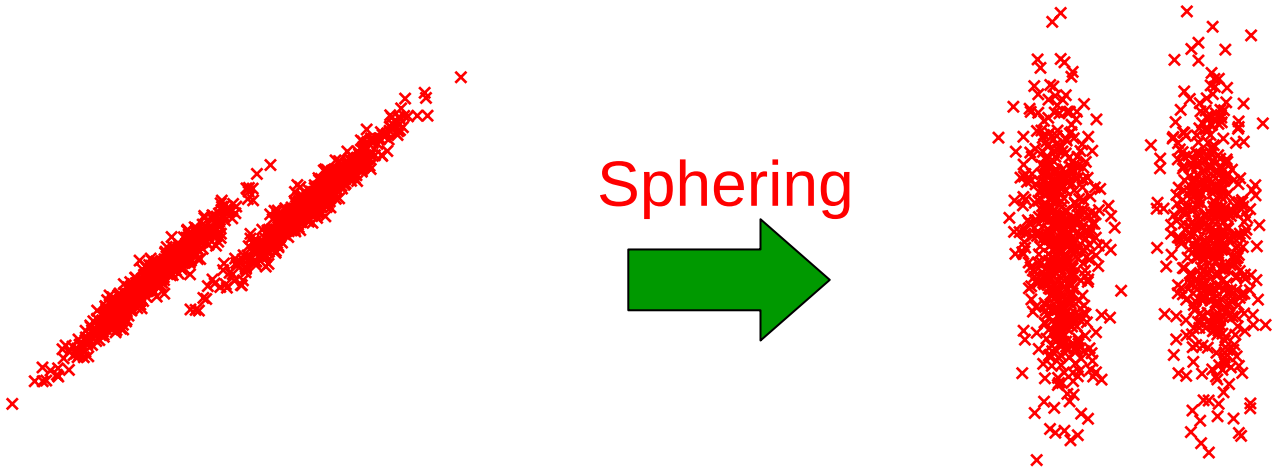
$$\boldsymbol{x}_i \longleftarrow \boldsymbol{x}_i - \overline{\boldsymbol{x}} \qquad \overline{\boldsymbol{x}} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x}_j$$

- Sphering (or pre-whitening):

$$\boldsymbol{x}_i \longleftarrow \boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{x}_i \qquad \boldsymbol{C} = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x}_j \boldsymbol{x}_j^\top$$

- Covariance matrix for sphered data:

$$\boldsymbol{C} = \boldsymbol{I}_d$$



Gradient for Sphered Data

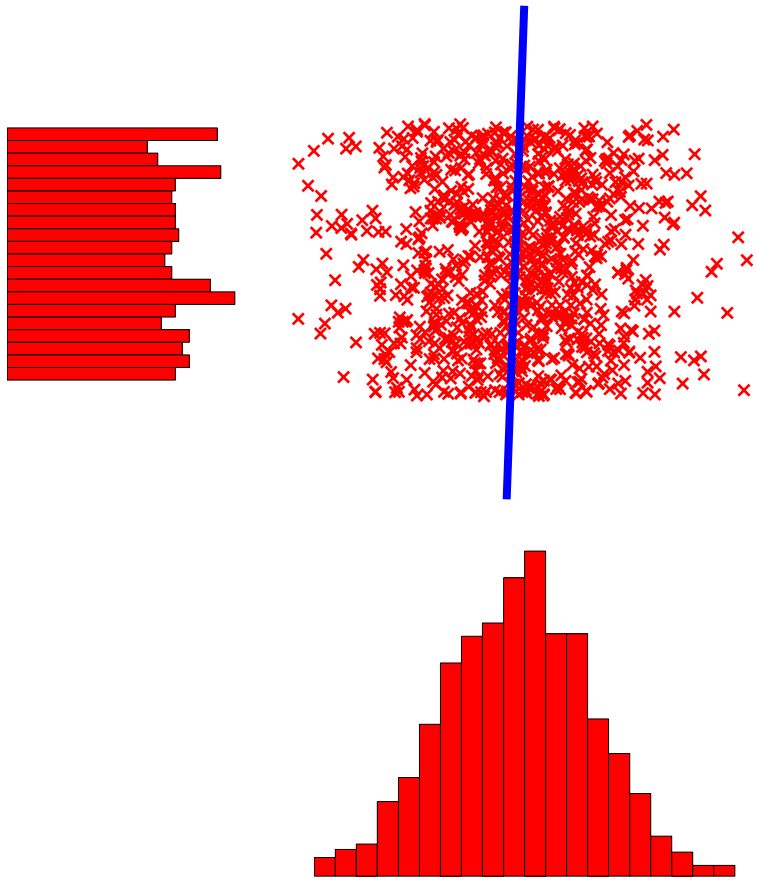
- For centered and sphered data, gradient is given by

$$\frac{\partial J_{PP}}{\partial \boldsymbol{b}} = 2\left(\frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{b}, \boldsymbol{x}_i\rangle^4 - 3\right)\left(\frac{4}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\langle \boldsymbol{b}, \boldsymbol{x}_i\rangle^3\right)$$

- Updating rule is
  - $\boldsymbol{b} \longleftarrow \boldsymbol{b} + \varepsilon\left(\frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{b}, \boldsymbol{x}_i\rangle^4 - 3\right)\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\langle \boldsymbol{b}, \boldsymbol{x}_i\rangle^3$
  - $\boldsymbol{b} \longleftarrow \boldsymbol{b} - \sum_{i=1}^{k-1}\langle \boldsymbol{b}, \boldsymbol{\psi}_i\rangle\boldsymbol{\psi}_i$
  - $\boldsymbol{b} \longleftarrow \boldsymbol{b}/\|\boldsymbol{b}\|$

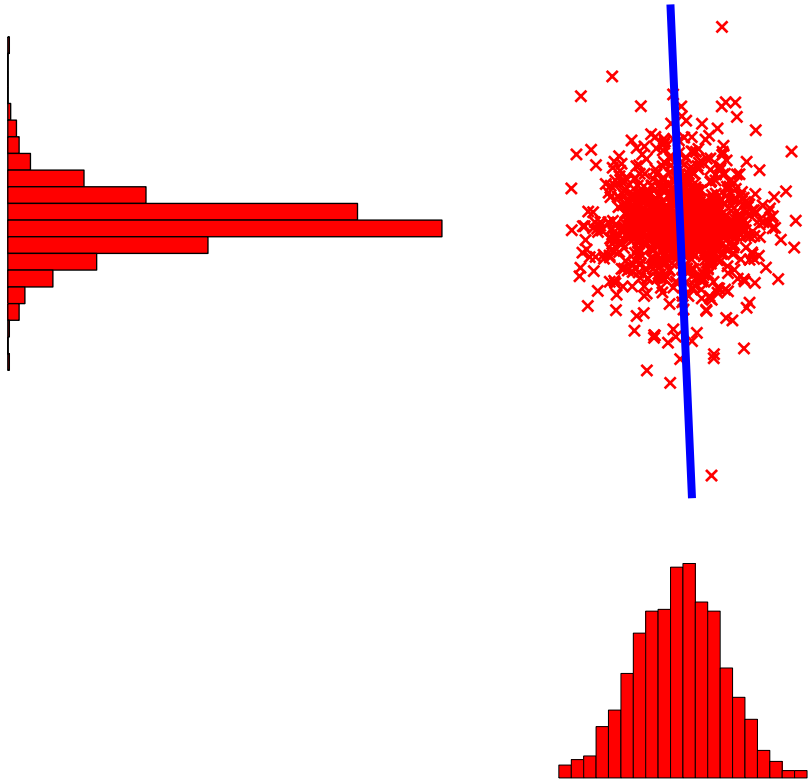
Examples

- $d = 2, \quad m = 1, \quad n = 1000$
- $\boldsymbol{x} = \begin{pmatrix} s \\ t \end{pmatrix}$
- $s \sim U(-\sqrt{3}, \sqrt{3})$
- $t \sim N(0, 1)$



Examples (cont.)

- $d = 2, \quad m = 1, \quad n = 1000$
- $\boldsymbol{x} = \begin{pmatrix} s \\ t \end{pmatrix}$
- $s \sim L(0, 1)$
- $t \sim N(0, 1)$



Homework

- Prove the followings for centered and sphered data:
  - Covariance matrix is given by

$$\frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{x}_j-\overline{\boldsymbol{x}})(\boldsymbol{x}_j-\overline{\boldsymbol{x}})^\top=\boldsymbol{I}_d$$

  - J_{PP} under $\|\boldsymbol{b}\|=1$ is given by

$$J_{PP}(\boldsymbol{b})=\left(\frac{1}{n}\sum_{i=1}^{n}\langle\boldsymbol{b},\boldsymbol{x}_i\rangle^4-3\right)^2$$

  - Gradient $\partial J_{PP}/\partial\boldsymbol{b}$ is given by

$$\frac{\partial J_{PP}}{\partial\boldsymbol{b}}=2\left(\frac{1}{n}\sum_{i=1}^{n}\langle\boldsymbol{b},\boldsymbol{x}_i\rangle^4-3\right)\left(\frac{4}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\langle\boldsymbol{b},\boldsymbol{x}_i\rangle^3\right)$$