Recap

- (Ch 1-2) Looking at data and relationships

Today

- (Ch 10) Data in high dimensions
  - Visualizing data
  - Summarizing data
  - Dimensionality reduction

Visualizing multidimensional data

- Visualizations (see today's Jupyter notebook)
  - 3D scatter plot (for 3-dimensional data)
  - Scatter plot matrix for 3 or more dimensions

- Reducing the dimensionality before visualization

Summarizing multidimensional data

- We need location and spread parameters for multidimensional data

- Notation
  - Suppose the dataset $\{\mathbf{x}\}$ consists of N items
  - Each item $\mathbf{x}_i$ is a d-dimensional vector of numbers
  - We refer to the jth component of the ith item as $\mathbf{x}_i^{(j)}$

Mean of a multidimensional dataset

- We compute the mean of $\{\mathbf{x}\}$ by computing the means of each component separately and stacking them into a vector

$$\text{mean of } j\text{th component} = \frac{\sum_i \mathbf{x}_i^{(j)}}{N}$$

- We write the mean of $\{\mathbf{x}\}$ as

$$\text{mean}(\{\mathbf{x}\}) = \frac{\sum_i \mathbf{x}_i}{N}$$

Covariance of a pair of components

- Recall that the covariance of random variables X and Y is

$$\mathrm{cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

- For a dataset with a pair of components $\{\mathbf{x}^{(j)}\}$ and $\{\mathbf{x}^{(k)}\}$

$$\mathrm{cov}\left(\{\mathbf{x}^{(j)}\}, \{\mathbf{x}^{(k)}\}\right) = \frac{\sum_i (\mathbf{x}_i^{(j)} - \mathrm{mean}(\{\mathbf{x}^{(j)}\}))(\mathbf{x}_i^{(k)} - \mathrm{mean}(\{\mathbf{x}^{(k)}\}))}{N}$$

Properties of covariance

- The covariance of a component with itself is its variance

$$\mathrm{cov}\left(\{\mathbf{x}^{(j)}\},\{\mathbf{x}^{(j)}\}\right) = \mathrm{var}\left(\{\mathbf{x}^{(j)}\}\right) = \mathrm{std}\left(\{\mathbf{x}^{(j)}\}\right)^2$$

- The correlation coefficient is the covariance scaled by standard deviations of each component

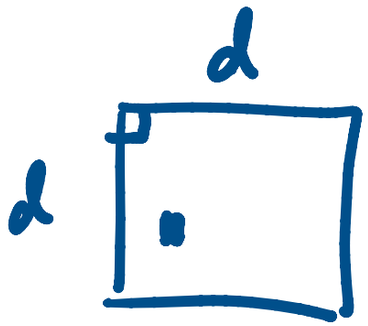
$$\mathrm{corr}\left(\{\mathbf{x}^{(j)},\mathbf{x}^{(k)}\}\right) = \frac{\sum_i \widehat{\mathbf{x}_i^{(j)}}\widehat{\mathbf{x}_i^{(k)}}}{N} = \frac{\mathrm{cov}\left(\{\mathbf{x}^{(j)}\},\{\mathbf{x}^{(k)}\}\right)}{\mathrm{std}(\{\mathbf{x}^{(j)}\})\mathrm{std}(\{\mathbf{x}^{(k)}\})}$$

Covariance matrix of multidimensional data

- We can capture all the pairwise covariances in a $d \times d$ matrix

$$\text{Covmat}(\{\mathbf{x}\}) = \frac{\sum_i (\mathbf{x}_i - \text{mean}(\{\mathbf{x}\}))(\mathbf{x}_i - \text{mean}(\{\mathbf{x}\}))^T}{N}$$

- Properties
  - The (j, k) entry of $\text{Covmat}(\{\mathbf{x}\})$ is $\text{cov}(\{\mathbf{x}^{(j)}\}, \{\mathbf{x}^{(k)}\})$
  - The (j, j) entry of $\text{Covmat}(\{\mathbf{x}\})$ is $\text{var}(\{\mathbf{x}^{(j)}\})$
  - $\text{Covmat}(\{\mathbf{x}\})$ is symmetric



Translation properties

- Translating the data translates the mean

$$\text{mean}(\{\mathbf{x} + \mathbf{c}\}) = \text{mean}(\{\mathbf{x}\}) + \mathbf{c}$$

- Translating the data leaves the covariance matrix unchanged

$$\text{Covmat}(\{\mathbf{x} + \mathbf{c}\}) = \text{Covmat}(\{\mathbf{x}\})$$

Proof

$$\text{Covmat}(\{\mathbf{x}+\mathbf{c}\})$$

$$= \frac{\sum_i (\mathbf{x}_i+\mathbf{c}-\text{mean}(\{\mathbf{x}+\mathbf{c}\}))(\mathbf{x}_i+\mathbf{c}-\text{mean}(\{\mathbf{x}+\mathbf{c}\}))^T}{N}$$

$$= \frac{\sum_i (\mathbf{x}_i+\mathbf{c}-\text{mean}(\{\mathbf{x}\})-\mathbf{c})(\mathbf{x}_i+\mathbf{c}-\text{mean}(\{\mathbf{x}\})-\mathbf{c})^T}{N}$$

by mean property

$$= \frac{\sum_i (\mathbf{x}_i-\text{mean}(\{\mathbf{x}\}))(\mathbf{x}_i-\text{mean}(\{\mathbf{x}\}))^T}{N}$$

$$= \text{Covmat}(\{\mathbf{x}\})$$

Linear transformation properties

- Linearly transforming the data linearly transforms the mean

$$\text{mean}(\{A\mathbf{x}\}) = A\ \text{mean}(\{\mathbf{x}\})$$

- Linearly transforming the data changes the covariance matrix

$$\text{Covmat}(\{A\mathbf{x}\}) = A\ \text{Covmat}(\{\mathbf{x}\})\ A^{\boldsymbol{T}}$$

$\text{var}(\{kx\}) = k^2\ \text{var}(\{x\})$

Proof

$$\text{Covmat}(\{A\mathbf{x}\})$$

$$= \frac{\sum_i (A\mathbf{x}_i - \text{mean}(\{A\mathbf{x}\}))(A\mathbf{x}_i - \text{mean}(\{A\mathbf{x}\}))^T}{N}$$

$$= \frac{\sum_i (A\mathbf{x}_i - A\,\text{mean}(\{\mathbf{x}\}))(A\mathbf{x}_i - A\,\text{mean}(\{\mathbf{x}\}))^T}{N}$$

by mean property

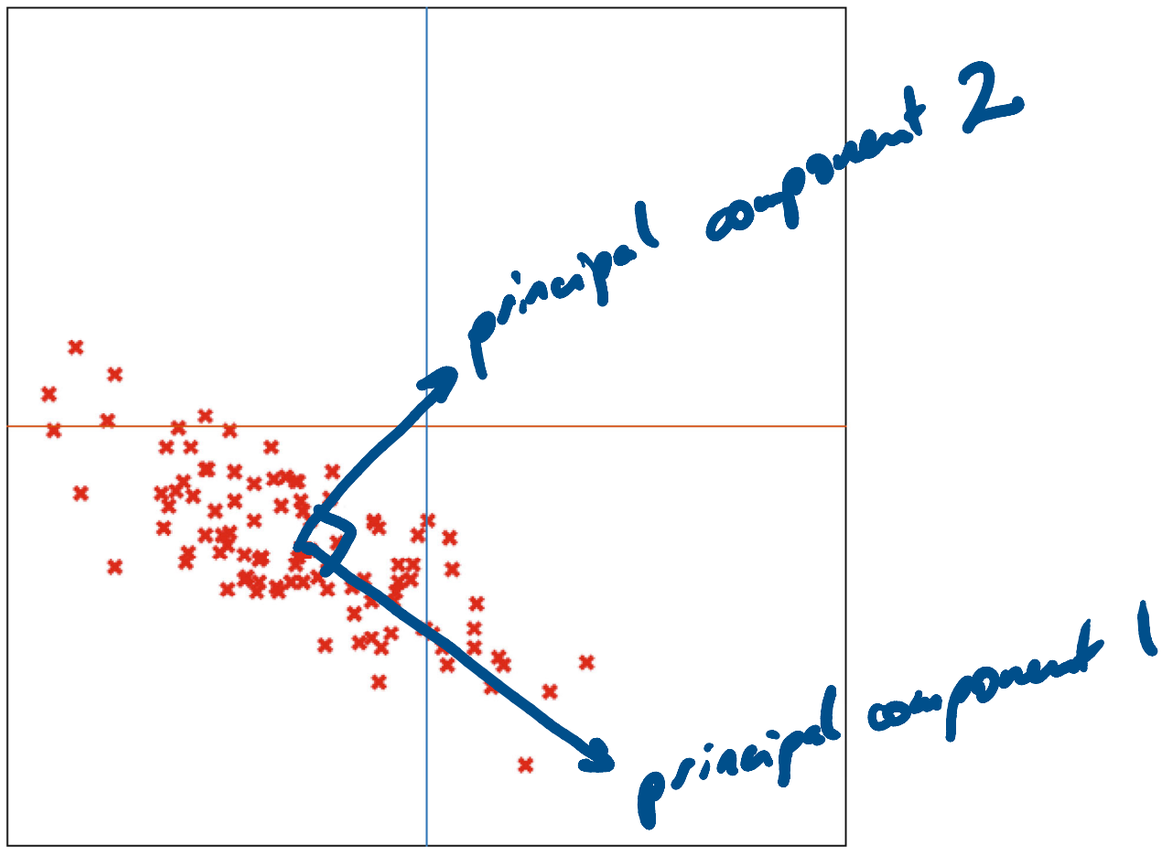
$(AB)^T = B^T A^T$

$$= \frac{\sum_i A(\mathbf{x}_i - \text{mean}(\{\mathbf{x}\}))(\mathbf{x}_i - \text{mean}(\{\mathbf{x}\}))^T A^T}{N}$$

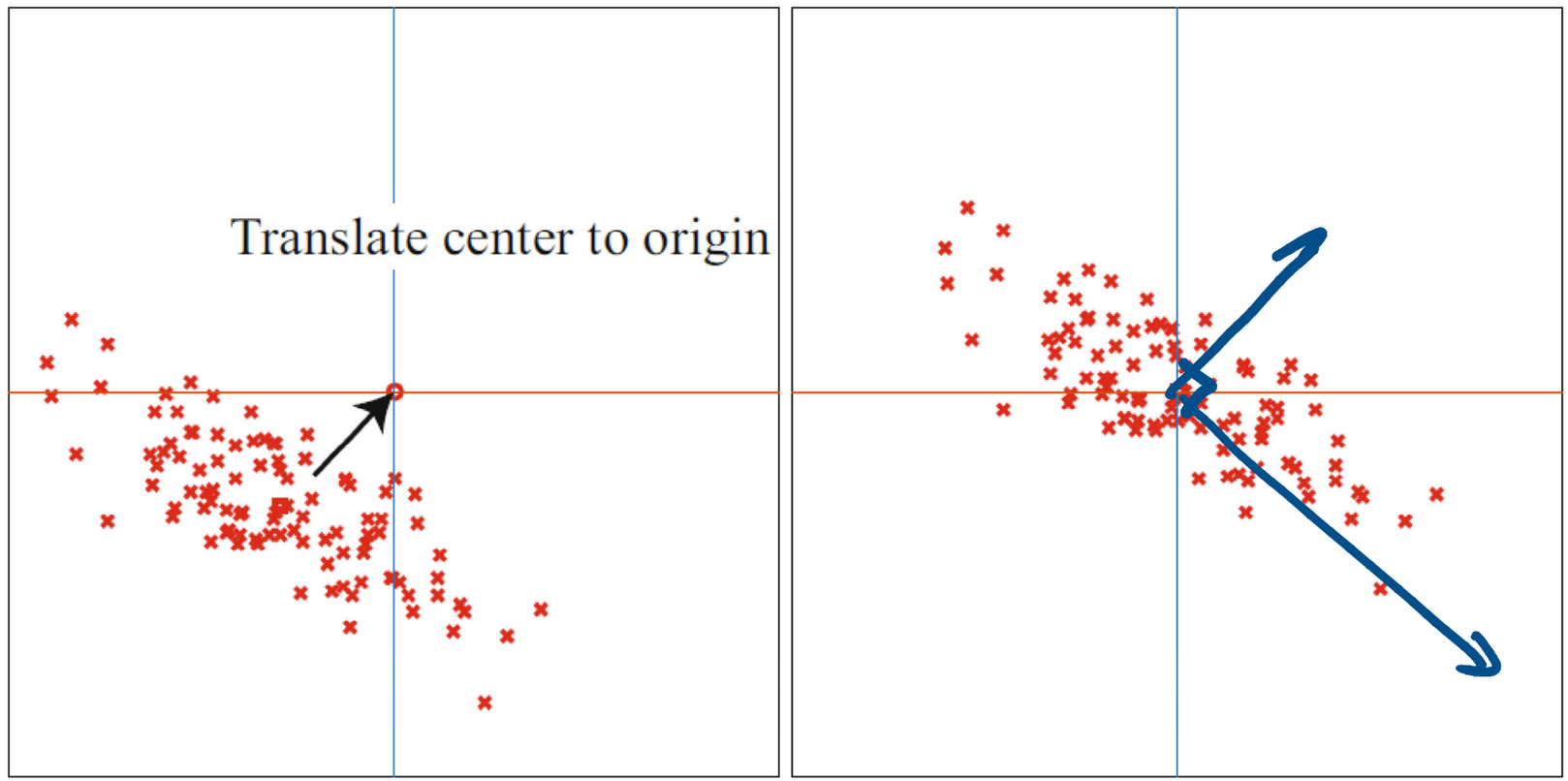
$$= A \frac{\sum_i (\mathbf{x}_i - \text{mean}(\{\mathbf{x}\}))(\mathbf{x}_i - \text{mean}(\{\mathbf{x}\}))^T}{N} A^T$$

$$= A\ \text{Covmat}(\{\mathbf{x}\})\ A^T$$

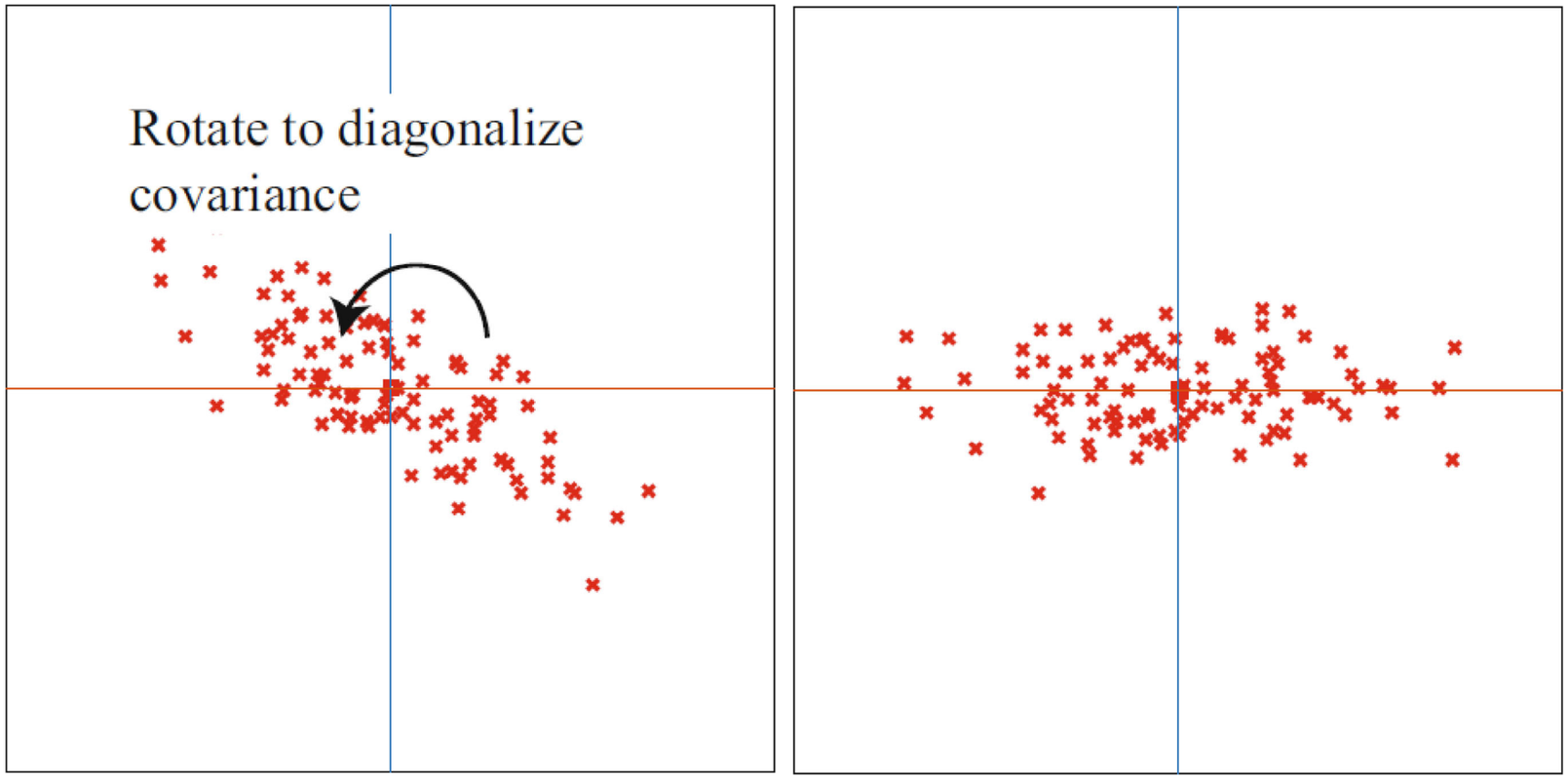
Dimensionality reduction: 2D to 1D example



Step 1: subtract mean



Step 2: apply linear transformation



Step 3: drop component(s)