# Announcements

# Principal Component Analysis

Machine Learning – CSE546
Kevin Jamieson
University of Washington

November 8, 2018

# Linear projections

Given $x_1, \ldots, x_n \in \mathbb{R}^d$, for $q \ll d$ find a compressed representation with $\lambda_1, \ldots, \lambda_n \in \mathbb{R}^q$ such that $x_i \approx \mu + \mathbf{V}_q \lambda_i$ and $\mathbf{V}_q^T \mathbf{V}_q = I$

$$\min_{\mu, \mathbf{V}_q, \{\lambda_i\}_i} \sum_{i=1}^n \|x_i - \mu - \mathbf{V}_q \lambda_i\|_2^2$$

# Linear projections

Given $x_1, \ldots, x_n \in \mathbb{R}^d$, for $q \ll d$ find a compressed representation with $\lambda_1, \ldots, \lambda_n \in \mathbb{R}^q$ such that $x_i \approx \mu + \mathbf{V}_q \lambda_i$ and $\mathbf{V}_q^T \mathbf{V}_q = I$

$$\min_{\mu, \mathbf{V}_q, \{\lambda_i\}_i} \sum_{i=1}^{n} \|x_i - \mu - \mathbf{V}_q \lambda_i\|_2^2$$

Fix $\mathbf{V}_q$ and solve for $\mu, \lambda_i$:

$$\mu = \bar{x} = \tfrac{1}{n} \textstyle\sum_{i=1}^n x_i$$

$$\lambda_i = \mathbf{V}_q^T (x_i - \bar{x})$$

Which gives us:

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||^2.$$

$\mathbf{V}_q \mathbf{V}_q^T$ is a *projection matrix* that minimizes error in basis of size $q$

# Linear projections

$$\sum_{i=1}^{N} ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||_2^2$$

$$\Sigma := \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$

$$\mathbf{V}_q^T \mathbf{V}_q = I_q$$

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||_2^2 = \min_{\mathbf{V}_q} Tr(\Sigma) - Tr(\mathbf{V}_q^T \Sigma \mathbf{V}_q)$$

Eigenvalue decomposition of $\sum =$

# Linear projections

$$\sum_{i=1}^{N} ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||_2^2$$

$$\Sigma := \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$

$$\mathbf{V}_q^T \mathbf{V}_q = I_q$$

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||_2^2 = \min_{\mathbf{V}_q} Tr(\Sigma) - Tr(\mathbf{V}_q^T \Sigma \mathbf{V}_q)$$

Eigenvalue decomposition of $\sum$ =

$\mathbf{V}_q$ are the first $q$ eigenvectors of $\Sigma$

Minimize reconstruction error and capture the most variance in your data.

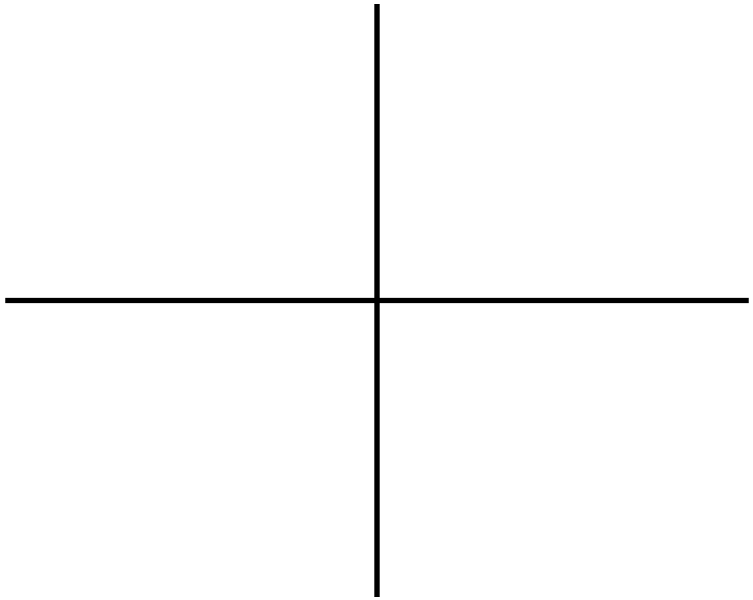
# Pictures

$\mathbf{V}_q$ are the first $q$ eigenvectors of $\Sigma$

$\mathbf{V}_q$ are the first q *principal components*

$$\Sigma := \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$

$$\mathbf{V}_q \text{ with } \mathbf{V}_q^T \mathbf{V}_q = I \text{ maximizes } Tr(\mathbf{V}_q^T \Sigma \mathbf{V}_q)$$



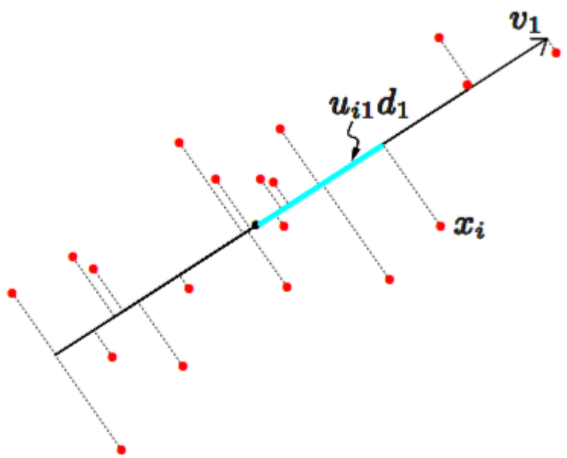
# Linear projections

Given $x_i \in \mathbb{R}^d$ and some $q < d$ consider

$$\min_{\mathbf{V}_q} \sum_{i=1}^{N} ||(x_i - \bar{x}) - \mathbf{V}_q \mathbf{V}_q^T (x_i - \bar{x})||^2.$$

where $\mathbf{V}_q = [v_1, v_2, \dots, v_q]$ is orthonormal:

$$\mathbf{V}_q^T \mathbf{V}_q = I_q$$




$\mathbf{V}_q$ are the first $q$ eigenvectors of $\Sigma$

$\mathbf{V}_q$ are the first q *principal components*

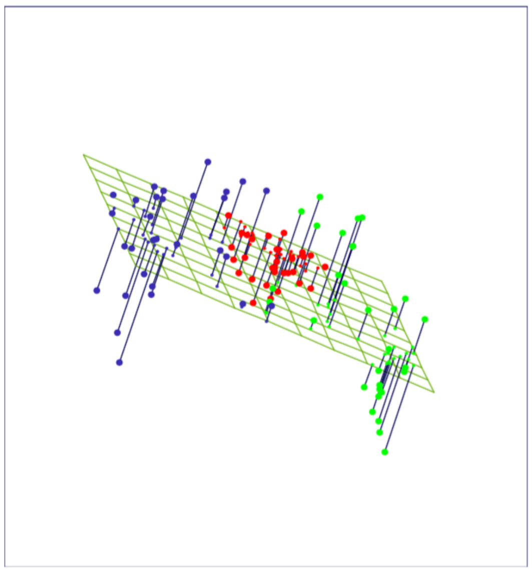
$$\Sigma := \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$

Principal Component Analysis (PCA) projects $(\mathbf{X} - \mathbf{1}\bar{x}^T)$ down onto $\mathbf{V}_q$

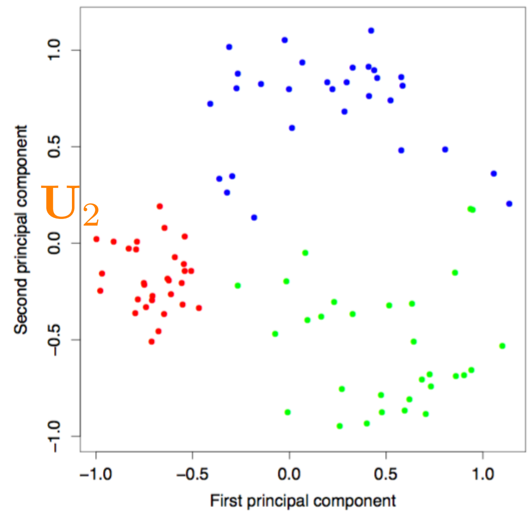
$$(\mathbf{X} - \mathbf{1}\bar{x}^T)\mathbf{V}_q = \mathbf{U}_q \mathrm{diag}(d_1, \dots, d_q) \qquad \mathbf{U}_q^T \mathbf{U}_q = I_q$$

# Dimensionality reduction

$\mathbf{V}_q$ are the first $q$ eigenvectors of $\Sigma$ and *SVD* $\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T$



$\mathbf{X} - \mathbf{1}\bar{x}^T$




$\mathbf{U}_2$

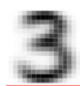
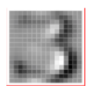
$\mathbf{U}_1$

# Dimensionality reduction

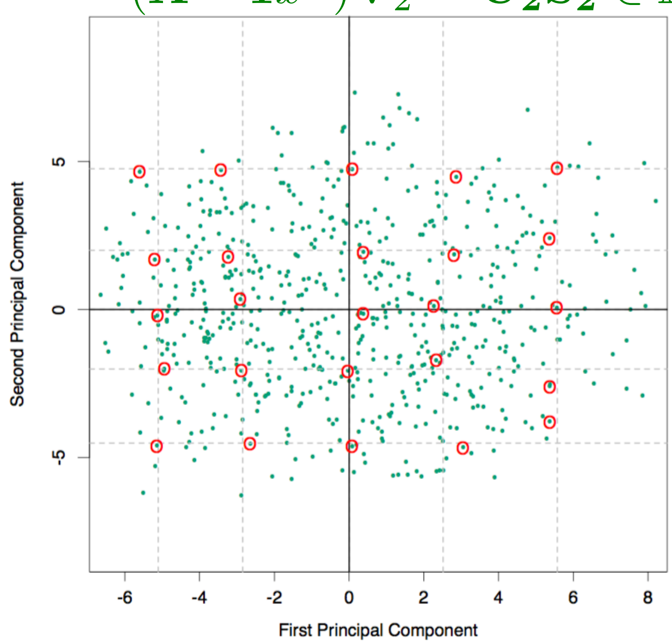
$\mathbf{V}_q$ are the first $q$ eigenvectors of $\Sigma$ and *SVD* $\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T$

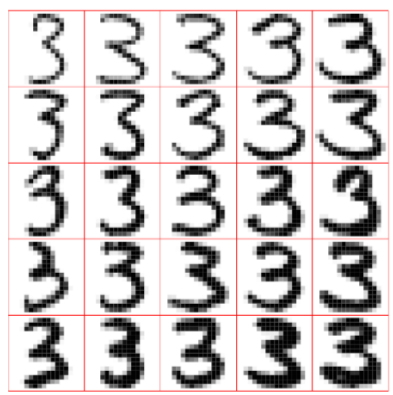
Handwritten 3's, 16x16 pixel image so that $x_i \in \mathbb{R}^{256}$

$$\hat{f}(\lambda) = \bar{x} + \lambda_1 v_1 + \lambda_2 v_2$$

$=$  $+ \lambda_1 \cdot$  $+ \lambda_2 \cdot$  .

$(\mathbf{X} - \mathbf{1}\bar{x}^T)\mathbf{V}_2 = \mathbf{U_2 S_2} \in \mathbb{R}^{n \times 2}$



diag($\mathbf{S}$)

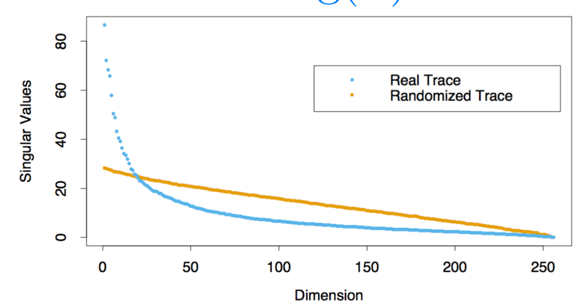



**FIGURE 14.24.** *The 256 singular values for the digitized threes, compared to those for a randomized version of the data (each column of* $\mathbf{X}$ *was scrambled).*



©Kevin Jamieson 2018

# Singular Value Decomposition (SVD)

**Theorem (SVD)**: Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rank $r \leq \min\{m, n\}$. Then $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ where $\mathbf{S} \in \mathbb{R}^{r \times r}$ is diagonal with positive entries, $\mathbf{U}^T\mathbf{U} = I$, $\mathbf{V}^T\mathbf{V} = I$.

$$\mathbf{A}^T\mathbf{A}v_i =$$

$$\mathbf{A}\mathbf{A}^T u_i =$$

# Singular Value Decomposition (SVD)

**Theorem (SVD)**: Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rank $r \leq \min\{m, n\}$. Then $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ where $\mathbf{S} \in \mathbb{R}^{r \times r}$ is diagonal with positive entries, $\mathbf{U}^T\mathbf{U} = I$, $\mathbf{V}^T\mathbf{V} = I$.

$$\mathbf{A}^T\mathbf{A}v_i = \mathbf{S}_{i,i}^2 v_i$$

$$\mathbf{A}\mathbf{A}^T u_i = \mathbf{S}_{i,i}^2 u_i$$

$\mathbf{V}$ are the first $r$ eigenvectors of $\mathbf{A}^T\mathbf{A}$ with eigenvalues $\text{diag}(\mathbf{S})$
$\mathbf{U}$ are the first $r$ eigenvectors of $\mathbf{A}\mathbf{A}^T$ with eigenvalues $\text{diag}(\mathbf{S})$

# Linear projections

$\mathbf{V}_q$ are the first $q$ eigenvectors of $\Sigma$

$\mathbf{V}_q$ are the first q *principal components*

$$\Sigma := \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$

Principal Component Analysis (PCA) projects $(\mathbf{X} - \mathbf{1}\bar{x}^T)$ down onto $\mathbf{V}_q$

$$(\mathbf{X} - \mathbf{1}\bar{x}^T)\mathbf{V}_q = \mathbf{U}_q \mathrm{diag}(d_1, \dots, d_q) \qquad \mathbf{U}_q^T \mathbf{U}_q = I_q$$

Singular Value Decomposition defined as

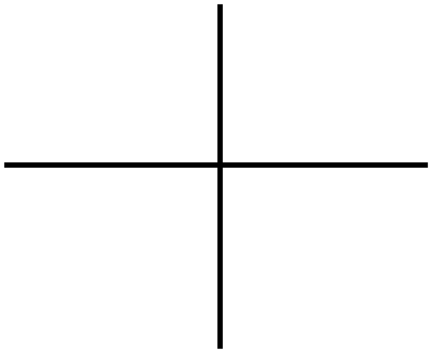
$$\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

# Pictures, intuition!

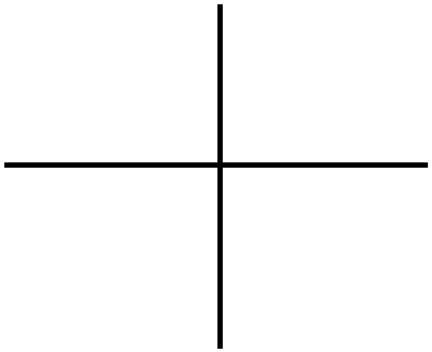
- Fill in the missing plots: $\mathbf{U}, \mathbf{S}, \mathbf{V} = \mathrm{svd}(\mathbf{JX})$

$$\mathbf{J} = I - \mathbf{1}\mathbf{1}^T/n \qquad \mathbf{JX} = \mathbf{USV}^T$$

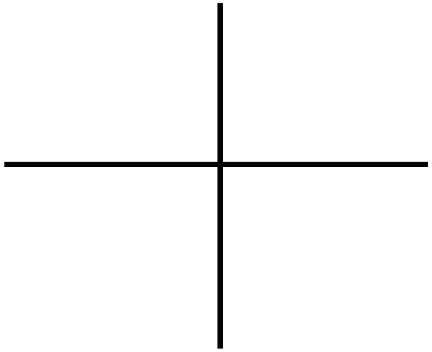
$\mathbf{X}$



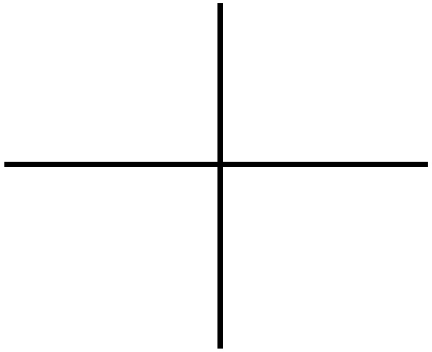
$\mathbf{JX}$



$\mathbf{JXVS}^{-1}$



$\mathbf{JXVS}^{-1}\mathbf{V}^T$



# Kernel PCA

$\mathbf{V}_q$ are the first $q$ eigenvectors of $\Sigma$ and *SVD* $\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{USV}^T$

$$(\mathbf{X} - \mathbf{1}\bar{x}^T)\mathbf{V}_q = \mathbf{U_q S_q} \in \mathbb{R}^{n \times q}$$

$$\mathbf{JX} = \mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{USV}^T \qquad \mathbf{J} = I - \mathbf{1}\mathbf{1}^T/n$$

$$(\mathbf{JX})(\mathbf{JX})^T =$$

# PCA Algorithm

**PCA**

**input**
A matrix of $m$ examples $X \in \mathbb{R}^{m,d}$
number of components $n$
**if** $(m > d)$
$A = X^\top X$
Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$ be the eigenvectors of $A$ with largest eigenvalues
**else**
$B = X X^\top$
Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be the eigenvectors of $B$ with largest eigenvalues
for $i = 1, \ldots, n$ set $\mathbf{u}_i = \frac{1}{\|X^\top \mathbf{v}_i\|} X^\top \mathbf{v}_i$
**output:** $\mathbf{u}_1, \ldots, \mathbf{u}_n$

# Cool tricks with SVD

Machine Learning – CSE546
Kevin Jamieson
University of Washington

November 8, 2016


©Kevin Jamieson 2018

# Ridge Regression revisited

$$\widehat{w}_{ridge} = \arg\min_w ||\mathbf{X}w - \mathbf{y}||_2^2 + \lambda||w||_2^2$$

$$\widehat{w}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

*Singular vector decomposition (SVD)*: $\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{USV}^T$

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

# Ridge Regression revisited

$$\widehat{w}_{ridge} = \arg\min_w ||\mathbf{X}w - \mathbf{y}||_2^2 + \lambda||w||_2^2$$

$$\widehat{w}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

*Singular vector decomposition (SVD)*: $\mathbf{X} - \mathbf{1}\bar{x}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T$

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y}$$

$$\hat{\mathbf{y}} = \sum_{i=1}^{d} u_i u_i^T \frac{s_i^2}{s_i^2 + \lambda} y_i$$

$$\mathbf{U} = [u_1, \ldots, u_d]$$

$$\mathbf{S} = \mathrm{diag}(s_1, \ldots, s_d)$$