Lecture 12

High Dimensional Geometry, SVD and Best-Fit Subspace

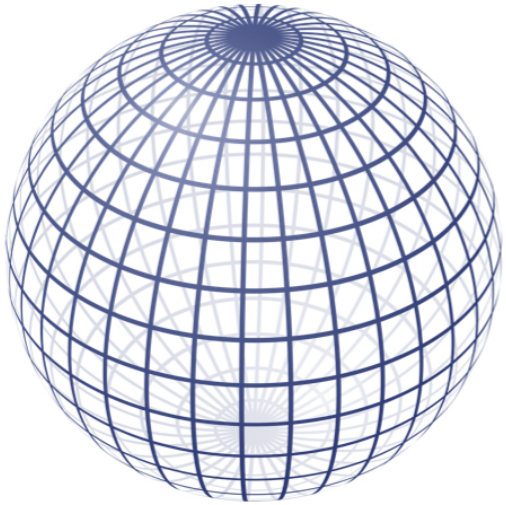
Bao-Jun Cai, 5/20/2026

Introduction to Algorithms for Data Science and Physics IMP@Fudan, 2026

Topics of this lecture:

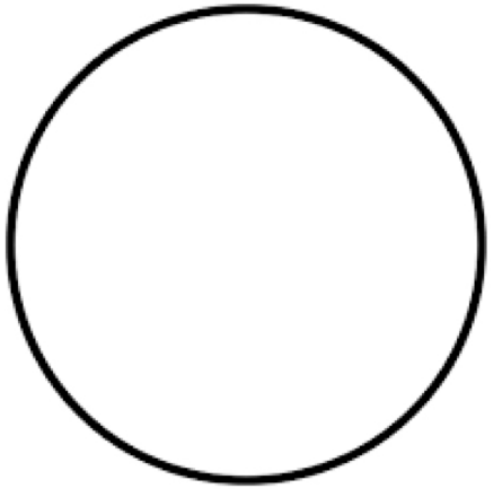
- volume of ball in high dimensions $\mathrm{Vol}(B((1-\epsilon)R))/\mathrm{Vol}(B(R)) = (1-\epsilon)^d$
- law of large numbers $P(|m^{-1}\sum_{i=1}^{m} x^{(i)} - \mathrm{E}[x]| \geq \epsilon) \leq \mathrm{var}[x]/m\epsilon^2$
- random projection theorem, Johnson-Lindernstrauss lemma
- singular-value decomposition $\mathbf{A} = \mathbf{U}\vec{\Sigma}\mathbf{V}^\top$
- largest variance, best-fit subspace, effective approximation $\mathbf{A} \approx \sum_{j=1}^{d_{\mathrm{eff}}} \sigma_j \mathbf{u}_j \mathbf{u}_j^\top$
- MNIST dataset "small" $\mathbf{y} \approx$ "large" $\mathbf{x}$

Circle, sphere, ...



volume: $V = 4\pi R^3/3$

surface: $S = 4\pi R^2$



volume: $V = \pi R^2$

surface: $S = 2\pi R$

volume: $V = 2R$

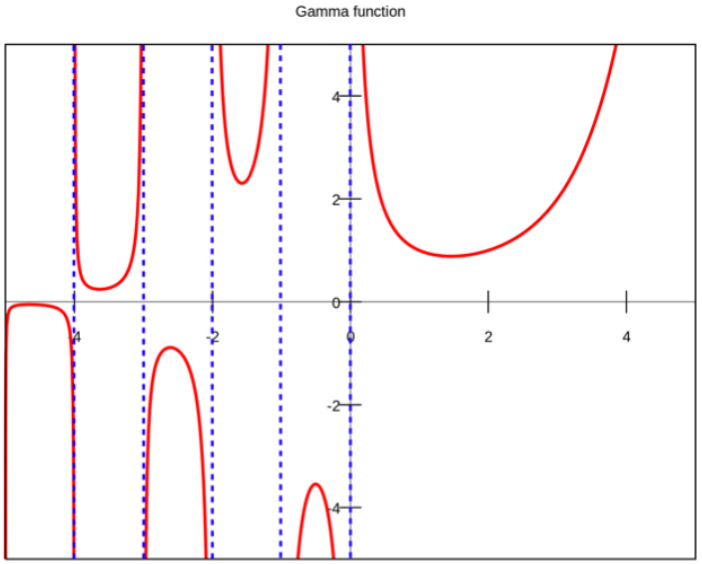
surface: $S = 2$

"Sphere" in d dimensions

Ex.: what is the expression for volume of the d-dimensional sphere?

$$x_1^2 + x_2^2 + \cdots + x_d^2 = R^2$$

$$\rightarrow \mathrm{Vol}(d) = \frac{2\pi^{d/2}R^d}{d\Gamma(d/2)} \sim R^d \rightarrow \mathrm{Suf}(d) = \frac{\partial \mathrm{Vol}(d)}{\partial R} = \frac{d}{R}\mathrm{Vol}(d)$$



$$\lim_{d\to\infty} \frac{\mathrm{Suf}(d)}{\mathrm{Vol}(d)} = \frac{d}{R} \rightarrow \infty$$

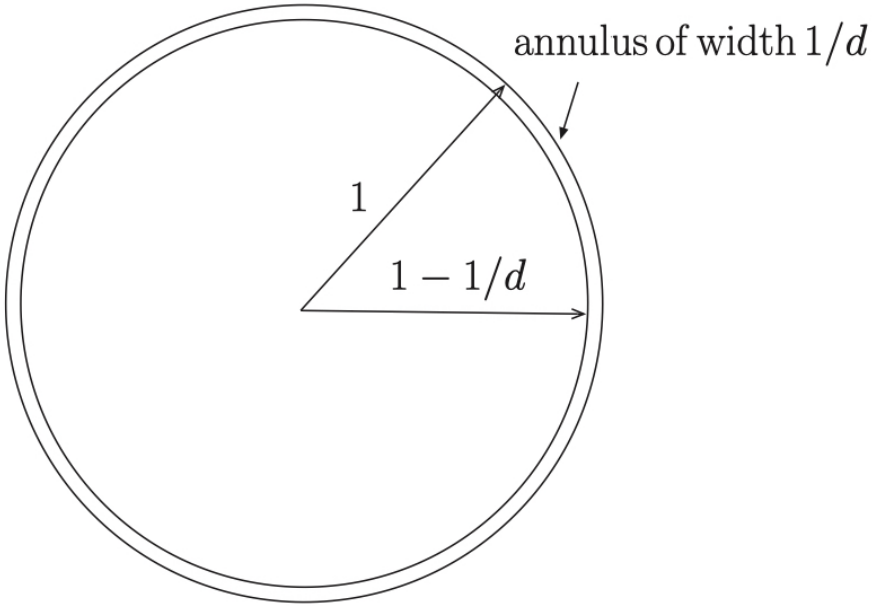
curse of dimensionality!

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$$

$$\Gamma(1/2) = \sqrt{\pi},\ \Gamma(1) = 1,\ \Gamma(3/2) = \sqrt{\pi}/2,\ \Gamma(2) = 1,\ \Gamma(5/2) = 3\sqrt{\pi}/4, \cdots$$

Theorem of annulus

$$B : \text{radius } R; \quad B' : \text{radius } R' = (1-\epsilon)R$$



$$\frac{\mathrm{Vol}(B')}{\mathrm{Vol}(B)} = (1-\epsilon)^d \le e^{-\epsilon d}$$

Ex.: prove that $(1-\epsilon)^d \le e^{-\epsilon d}$

Law of large numbers

$$\text{Prob}\left(\left|\frac{x^{(1)} + x^{(2)} + \cdots + x^{(m)}}{m} - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{m\epsilon^2}$$

$$x^{(i)} \sim p(x|\mu, \sigma^2)$$

proof:
Markov inequality
+Chebyshev inequality

Markov: $P(x \geq a) \leq \mathrm{E}[x]/a,\ x \geq 0$

Chebyshev: $P(|x - \mathrm{E}[x]| \geq c) \leq \mathrm{var}[x]/c^2$

Two random vectors in dimensions d

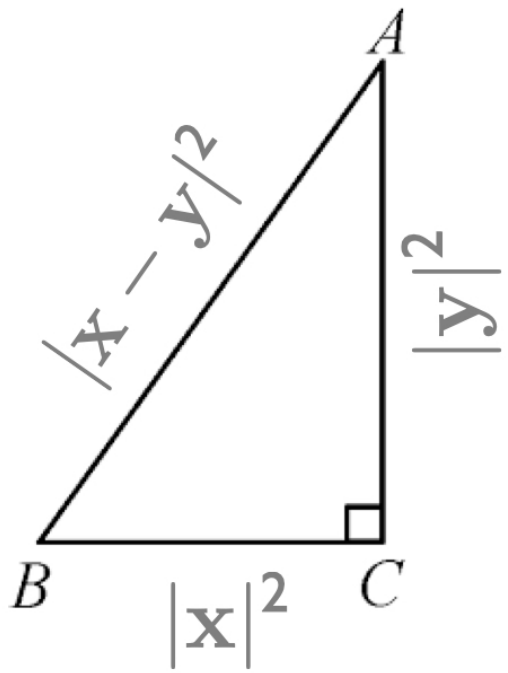
$$\mathbf{x}, \mathbf{y} \in \mathbf{R}^d, \; x_i, y_i \sim \mathcal{N}(0, 1)$$

$$E[\mathbf{x}^2] = \sum_{i=1}^{d} \mathrm{E}[x_i^2] = d\mathrm{var}[x_i] = d$$

$$\text{distance} = |\mathbf{x} - \mathbf{y}|^2 = \sum_{i=1}^{d} (x_i - y_i)^2$$

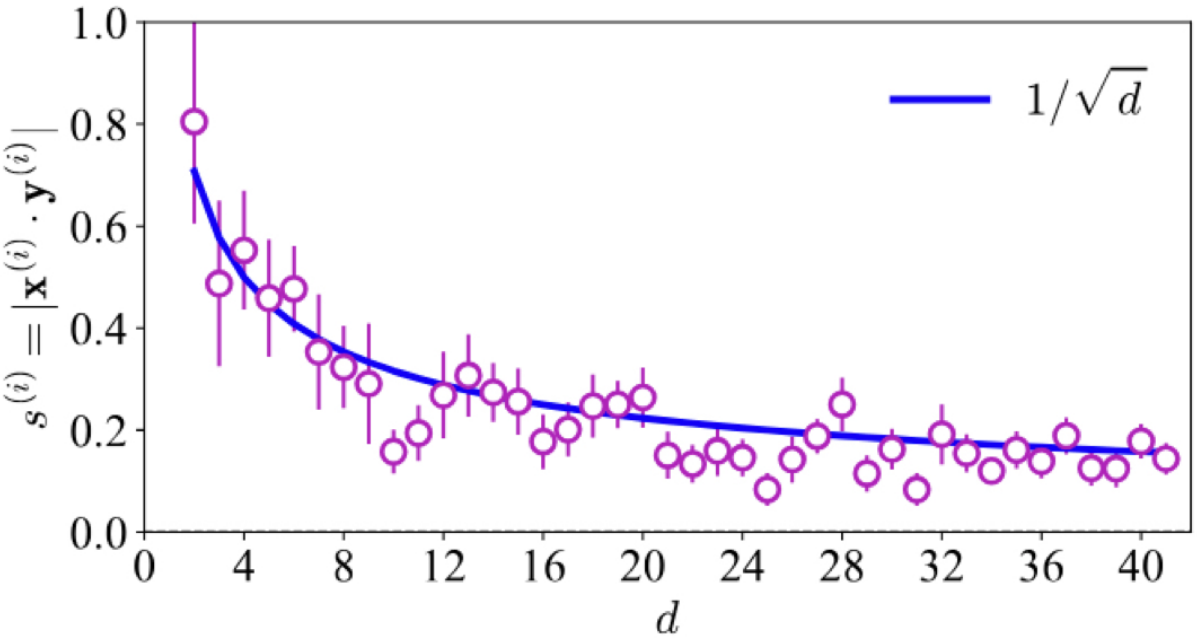
$$= \sum_{i=1}^{d} \left(\mathrm{E}[x_i^2] + \mathrm{E}[y_i^2] - 2\mathrm{E}[x_i]\mathrm{E}[y_i] \right)$$

$$= \sum_{i=1}^{d} \left(\mathrm{var}[x_i] + \mathrm{var}[y_i] - 2\mathrm{E}[x_i]\mathrm{E}[y_i] \right) = 2d$$



$$|\mathbf{x} - \mathbf{y}|^2 \approx |\mathbf{x}|^2 + |\mathbf{y}|^2$$

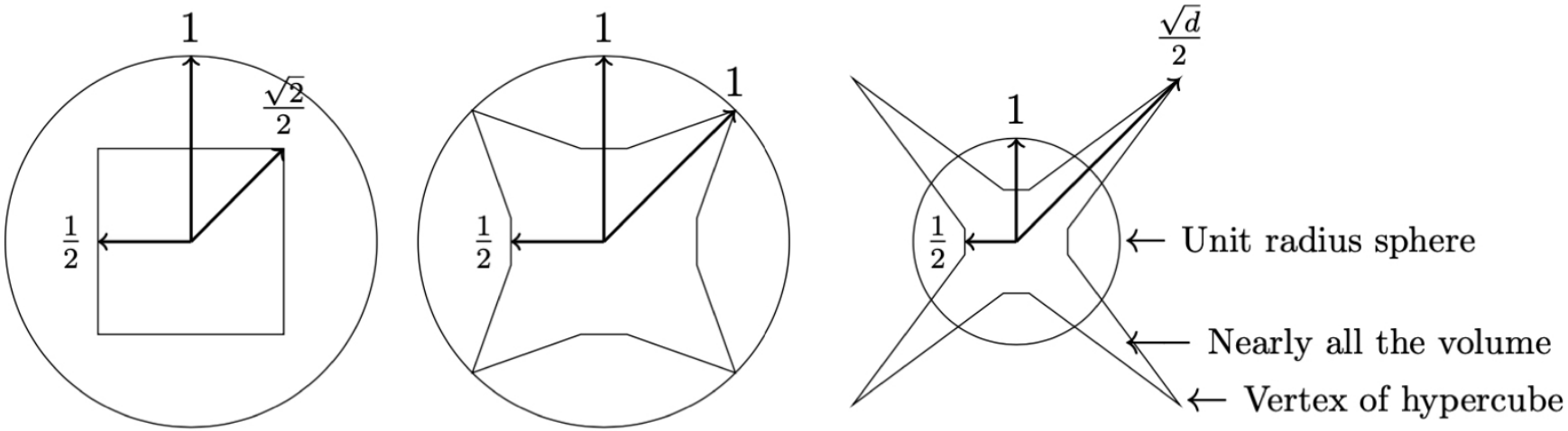
Simulation for vectors in high-D



$$\|\mathbf{x}^{(i)}\| \geq 1 - \frac{2 \ln m}{d}, \ \mathbf{x}^{(i),\top}\mathbf{x}^{(j)} = \mathbf{x}^{(i)} \cdot \mathbf{x}^{(j)} \leq \sqrt{\frac{6 \ln m}{d-1}}$$

How to understand the result?

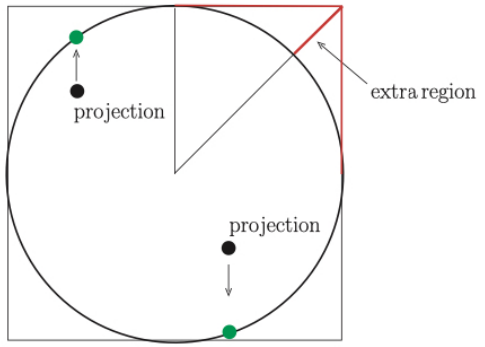
A. Blum, J. Hopcroft, and R. Kannan, *Foundations of Data Science*, Cambridge, 2020.



one can show that most of the volume of the unit ball in high dimensions is concentrated near its "equator", for for any unit-length vector **a** defining the "north pole", most of the volume of the unit ball lies in the thin slab of points whose inner product with **a** has magnitude O(d^{-1/2})

Generate uniform random samples on surface/volume

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{x_1^2 + x_2^2 + \cdots + x_d^2}{2}\right), \quad x_i \sim \text{Box-Muller}$$

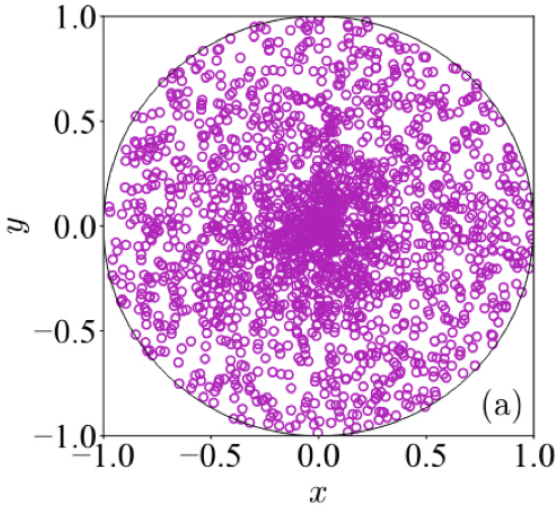


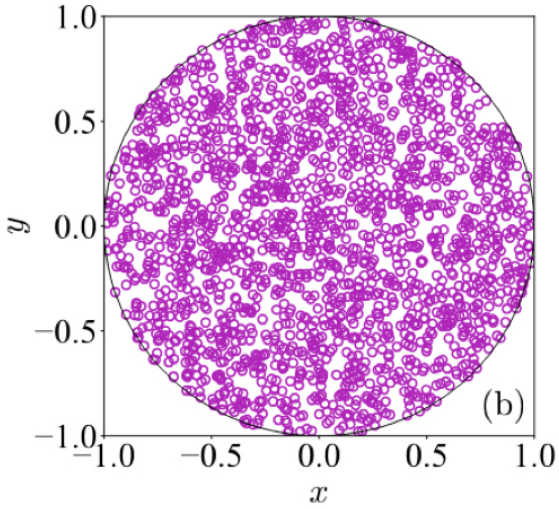
???

surface: $\mathbf{x}/|\mathbf{x}|$

$\rho(r) = cr^{d-1}$

volume: $\rho(r)\mathbf{x}/|\mathbf{x}|$





Random projection theorem

$$\mathbf{a}^{(i)} \in R^d,\ i = 1 \sim s,\ a_j^{(i)} \sim \mathcal{N}(0,1)$$

$$\mathbf{R}^d \to \mathbf{R}^s : \mathbf{f}(\mathbf{b}) = (\mathbf{a}^{(1)} \cdot \mathbf{b}, \cdots, \mathbf{a}^{(s)} \cdot \mathbf{b})^\top$$

$$|\mathbf{f}(\mathbf{b})| \approx \sqrt{s}|\mathbf{b}| : \text{Prob}\left(\left||\mathbf{f}(\mathbf{b})| - \sqrt{s}|\mathbf{b}|\right| \geq \epsilon\sqrt{s}|\mathbf{b}|\right) \leq 3e^{-\phi s\epsilon^2},\ \phi > 0$$

$$\text{Prob} \sim 1 - 2\binom{m}{2}e^{-s\epsilon^2/8},\ s \gtrsim \frac{16}{\epsilon^2}\ln\frac{m}{\epsilon}$$

$$(1-\epsilon)\sqrt{s}\left|\mathbf{a}^{(i)} - \mathbf{a}^{(j)}\right| \leq \left|\mathbf{f}(\mathbf{a}^{(i)}) - \mathbf{f}(\mathbf{a}^{(j)})\right| \leq (1+\epsilon)\sqrt{s}\left|\mathbf{a}^{(i)} - \mathbf{a}^{(j)}\right|$$

Johnson-Lindernstrauss lemma

Singular-value-decomposition (SVD)

d unknowns, *m* equations

$$\mathbf{Ax} = \mathbf{b},\ \mathbf{A} \in \mathbf{R}^{m \times d}$$

$$\begin{aligned}
a_1^{(1)} x_1 + a_2^{(1)} x_2 + \cdots + a_d^{(1)} x_d &= b_1 \\
a_1^{(2)} x_1 + a_2^{(2)} x_2 + \cdots + a_d^{(2)} x_d &= b_2 \\
&\vdots \\
a_1^{(m)} x_1 + a_2^{(m)} x_2 + \cdots + a_d^{(m)} x_d &= b_m
\end{aligned}$$

SVD of $\mathbf{A}$

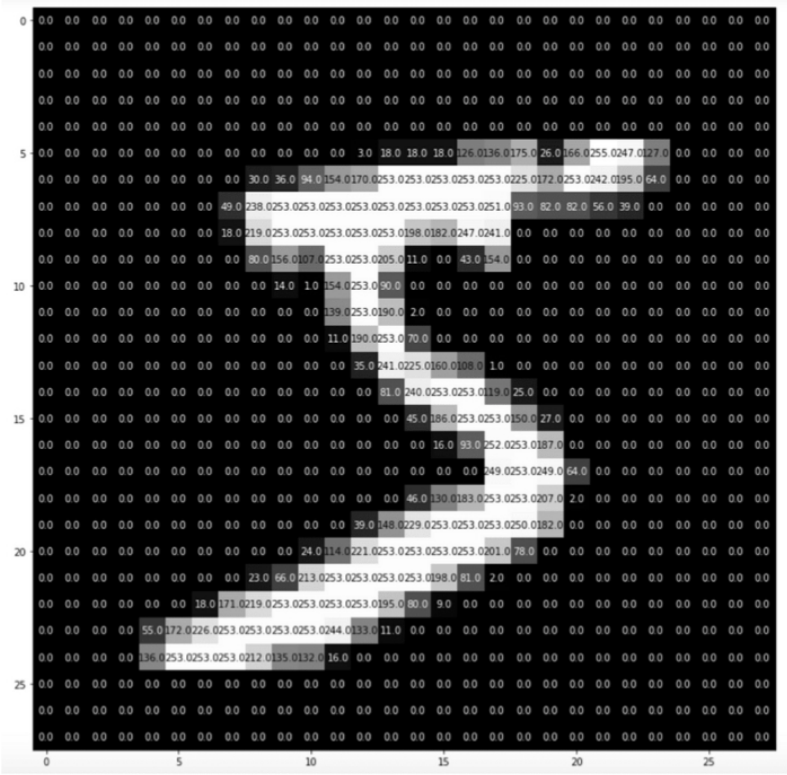
$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$$

$$\mathbf{A} = \mathbf{U}\vec{\Sigma}\mathbf{V}^\top$$

$$\mathbf{U} \in \mathbf{R}^{m \times d}, \vec{\Sigma} \in \mathbf{R}^{d \times d}, \mathbf{V} \in \mathbf{R}^{d \times d}$$

$$a_j^{(i)} = \sum_{k=1}^{d} \sigma_k U_{ik} V_{jk}$$

Another motivation for SVD



$$x \in R^{28\times 28} \sim R^{784} \sim R^{b}$$

"small" $y \approx x$?

Best-fit subspace

square: $\mathbf{A} = \sum_{j=1}^{d} \sigma_j \mathbf{u}_j \mathbf{u}_j^\top$

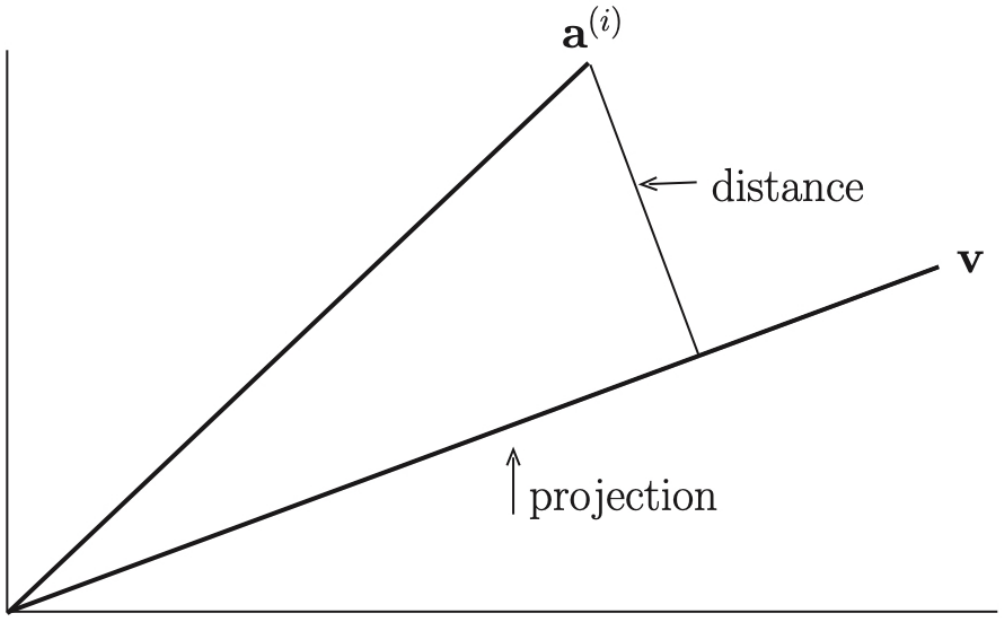
separable of SVs

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{d_{\text{eff}}} \gg \sigma_{d_{\text{eff}}+1} \geq \cdots \geq \sigma_d$$

$$\mathbf{A} \approx \sum_{j=1}^{d_{\text{eff}}} \sigma_j \mathbf{u}_j \mathbf{u}_j^\top$$

d_{eff} could be far smaller than d

Geometric meaning of best-fit subspace



$$\mathbf{v}_1 = \text{argmax}_{|\mathbf{v}|=1} |\mathbf{A}\mathbf{v}|^2$$

$$\mathbf{v}_2 = \text{argmax}_{\mathbf{v}\perp\mathbf{v}_1, |\mathbf{v}|=1} |\mathbf{A}\mathbf{v}|^2$$

$$\vdots$$

$$[\text{distance of point to line}]^2 = a_1^{(i),2} + a_2^{(i),2} + \cdots + a_d^{(i),2} - \overbrace{[\text{length of projection}]^2}^{|\mathbf{A}\mathbf{v}|^2}$$

Directions with largest variance

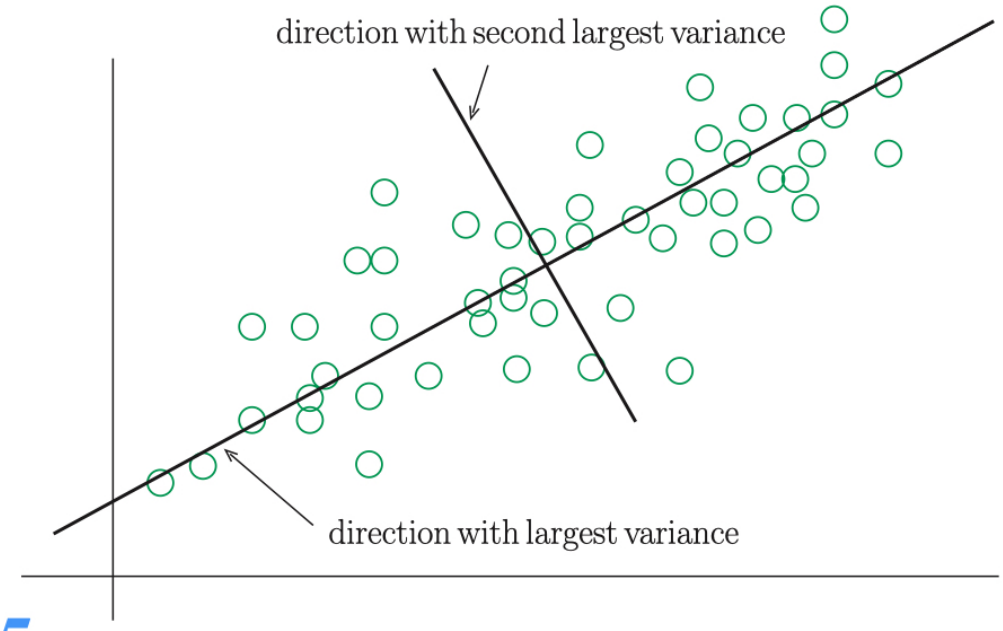
definition:

for $\mathbf{x} \in \mathbb{R}^b$, define the first s principle components y_j's such that $\mathrm{var}[y_1] \geq \mathrm{var}[y_2] \cdots \geq \mathrm{var}[y_s]$

$$y_1 = \mathbf{u}_1^\top \mathbf{x},\ \mathbf{u}_1^\top \mathbf{u}_1 = 1,\ \mathbf{u}_1 \in \mathbb{R}^b$$

$$\mathbf{u}_1^* = \mathrm{argmax}_{|\mathbf{u}|=1} \mathrm{var}[\mathbf{u}^\top \mathbf{x}]$$

Ex.: find out $\mathbf{u}_1$ and generally $\mathbf{u}_j$



Stacking principal components

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix} \in \mathbb{R}^s,\ \mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_s) \in \mathbb{R}^{b \times s}$$

$$\mathbf{y} = \mathbf{U}^\top \mathbf{x}$$

$$\vec{\Sigma}_{\mathbf{y}} = \mathbf{U}^\top \vec{\Sigma}_{\mathbf{x}} \mathbf{U}$$

$$\hat{\vec{\Sigma}}_m = \frac{1}{m} \sum_{i=1}^{m} \mathbf{x}^{(i)} \mathbf{x}^{(i),\top} = \frac{1}{m} \mathbf{X}\mathbf{X}^\top \approx \vec{\Sigma}_{\mathbf{x}}$$

An optimization problem for reconstruction

$$\min_{\mathbf{U}^\top \mathbf{U}=\mathbf{I}_s,\{\mathbf{y}^{(i)}\}} \sum_{i=1}^{m} \|\mathbf{x}^{(i)} - \mathbf{U}\mathbf{y}^{(i)}\|^2, \ \sum_{i=1}^{m} \mathbf{y}^{(i)} = \mathbf{0}$$

original data

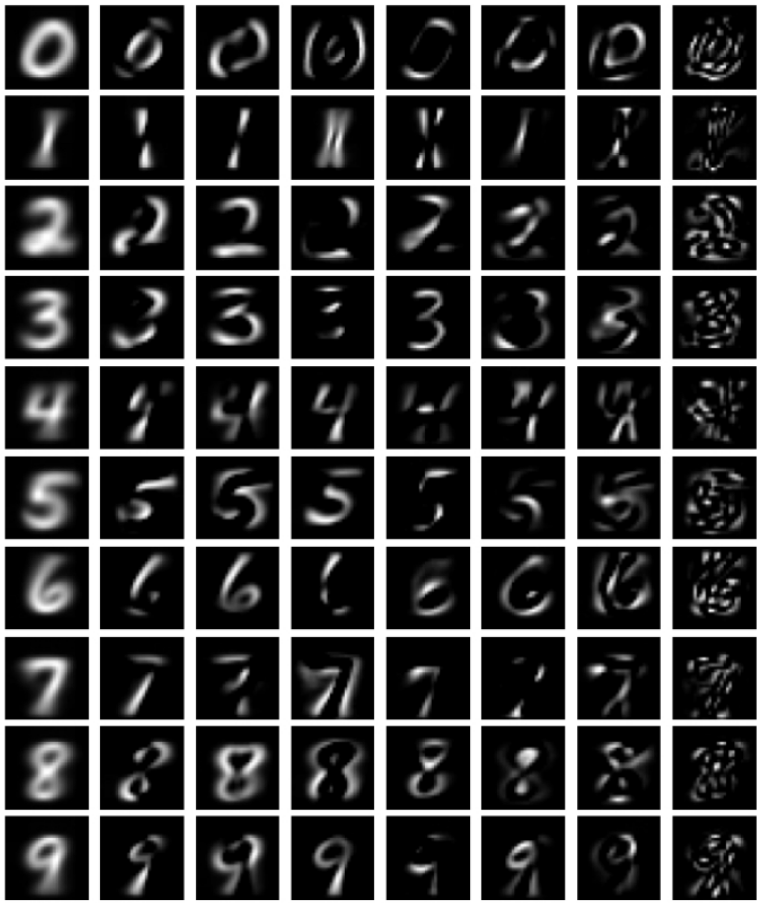
reconstruction of $\mathbf{x}^{(i)}$

$$\min_{\mathbf{U}^\top \mathbf{U}=\mathbf{I}_s} \sum_{i=1}^{m} \|\mathbf{x}^{(i)} - \mathbf{U}\mathbf{U}^\top \mathbf{x}^{(i)}\|^2$$

$$\text{error} \sum_{i=s+1}^{b} \sigma_i^2$$

Ex.: SVD of $\mathbf{X}\mathbf{X}^\top$ is necessary, why?

Notations for image application(s)



$$\mathbf{X}_{(k)} = \left(\mathbf{x}_{(k)}^{(1)}, \cdots, \mathbf{x}_{(k)}^{(m)}\right) \in \mathbf{R}^{b \times m}$$

$$\vec{\mu}_{(k)} = m^{-1} \sum_{i=1}^{m} \mathbf{x}_{(k)}^{(i)}$$

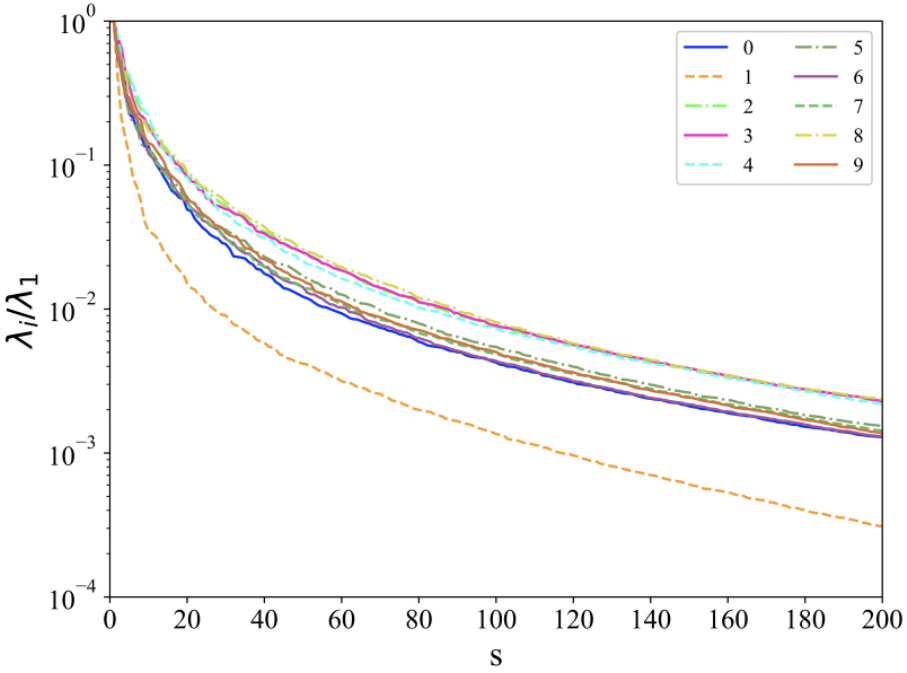
$$\widehat{\mathbf{X}}_{(k)} = \mathbf{X}_{(k)} - \vec{\mu}_{(k)}$$

$$\widehat{\mathbf{X}}_{(k)} \widehat{\mathbf{X}}_{(k)}^{\top} = \mathbf{U}_{(k)} \mathrm{diag}(\lambda_{1,(k)}, \cdots, \lambda_{784,(k)}) \mathbf{U}_{(k)}^{\top}$$

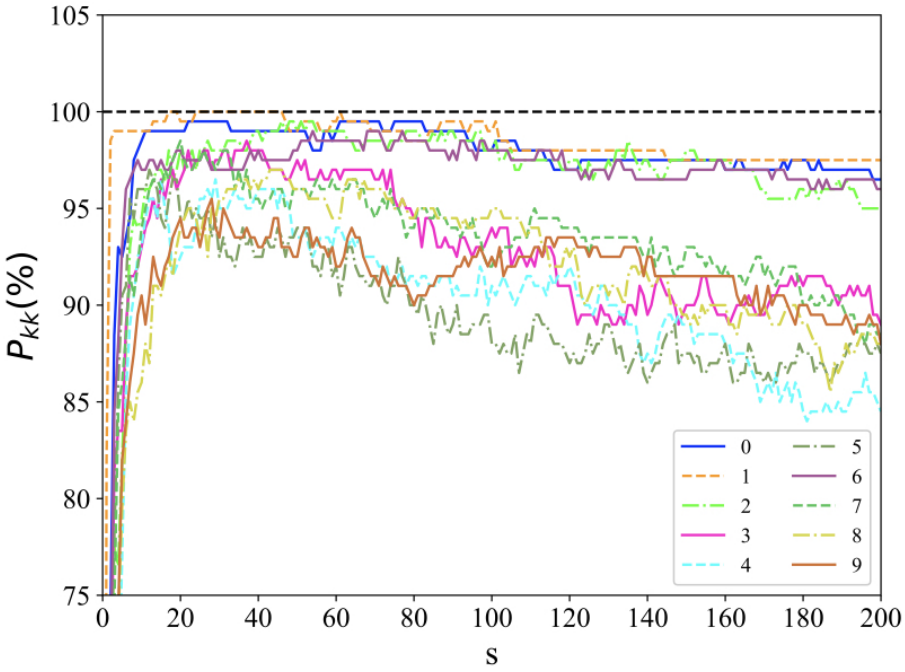
$$\Delta_{(k)} = \mathrm{tr}\left(\mathbf{U}_{(k)}^{(s),\top} \widehat{\mathbf{f}}_{(k)} \widehat{\mathbf{f}}_{(k)}^{\top} \mathbf{U}_{(k)}^{(s)}\right)$$

$$k^{*} = \mathrm{argmax}_{k} \Delta(k)$$

Sis and correctness

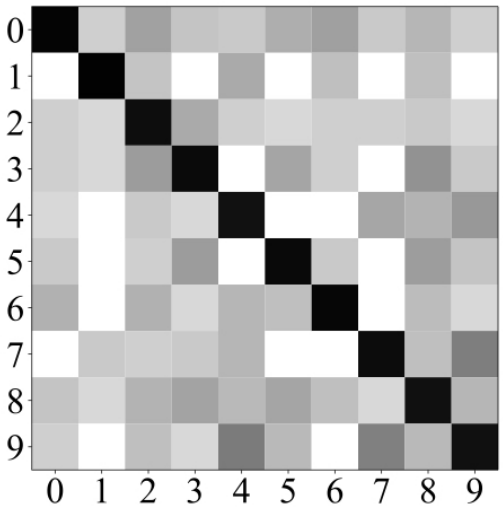


$$\lambda_i = \sigma_i^2$$

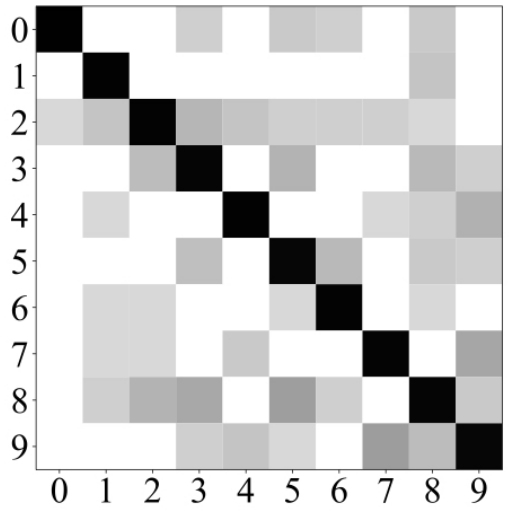


Correlation(s)

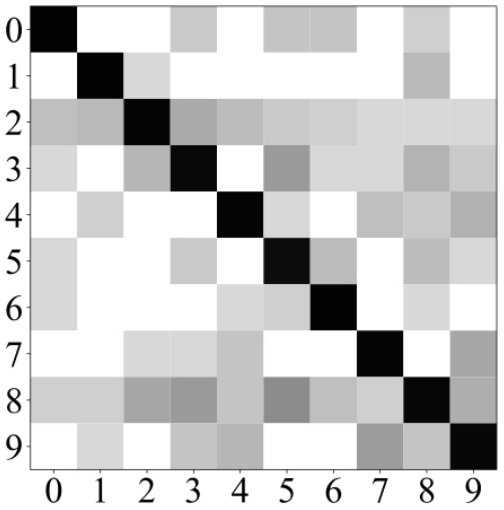
sum rule for each row $\sum_{k'=0}^{9} P_{kk'} = 1$



s=5

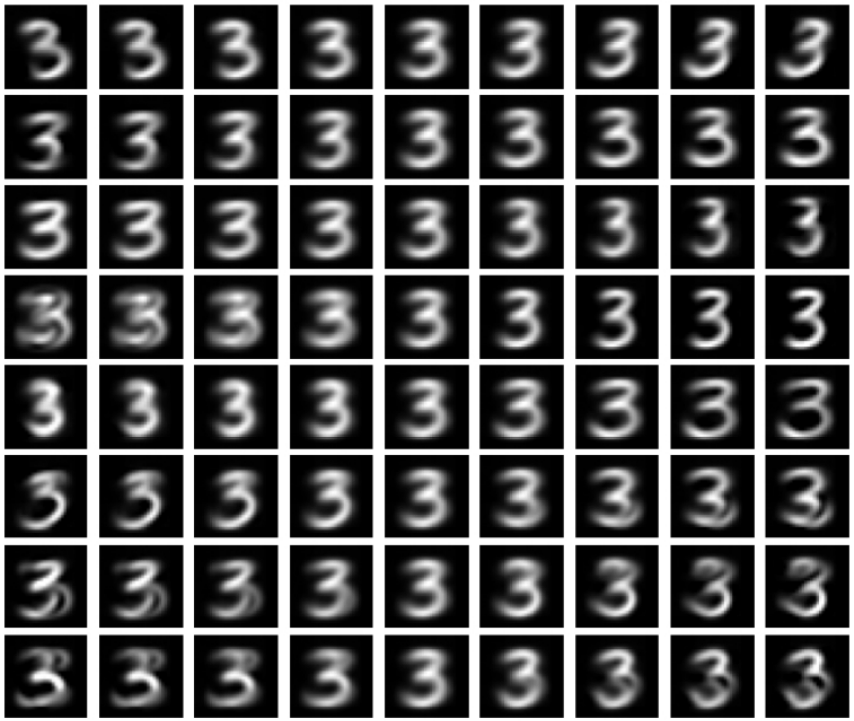


s=40

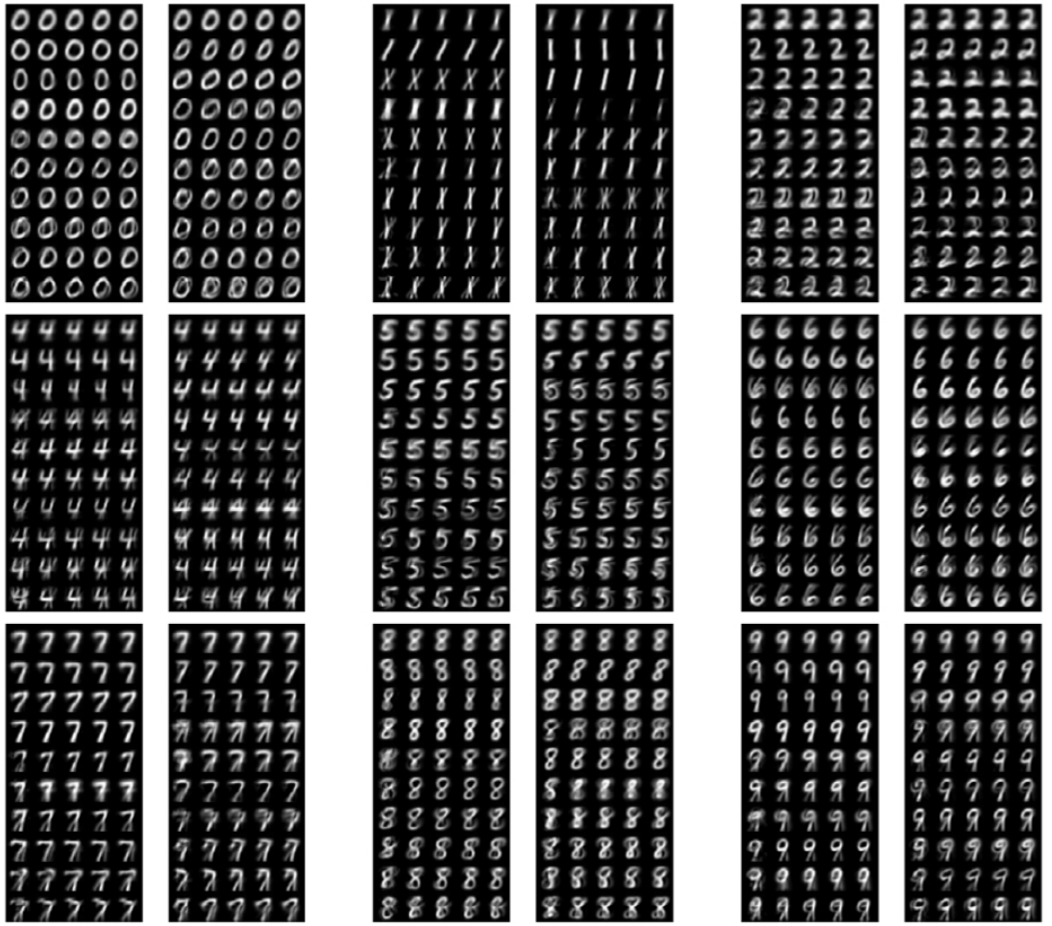


s=100

Feature(s) discovered by PCA

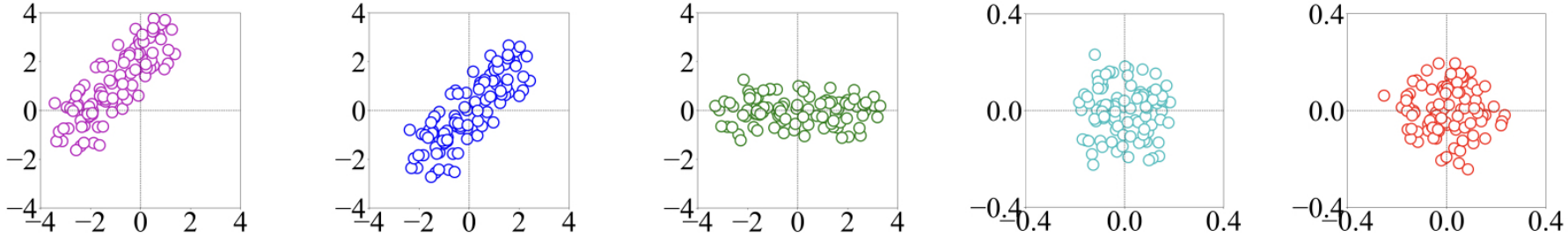


$$\hat{\vec{\mu}}_m + \chi \mathbf{u}_i,\ i = 1 \sim 8$$



Application: whitening & SPCA*

$$\mathbf{x}^{(i)} \leftarrow \mathbf{x}^{(i)} - \vec{\mu},\ \mathbf{x}^{(i)} \leftarrow \mathbf{V}^\top \mathbf{x}^{(i)},\ \mathbf{x}_j^{(i)} \leftarrow \mathbf{x}^{(i)} / \sigma_j,\ \mathbf{x}^{(i)} \leftarrow \mathbf{V}\mathbf{x}^{(i)}$$



sparse PCA (SPCA):

$$\widehat{\mathbf{B}}) = \text{argmin}_{\mathbf{A},\mathbf{B}} \left[\sum_{i=1}^{m} \left\| \mathbf{x}^{(i)} - \mathbf{A}\mathbf{B}^\top \mathbf{x}^{(i)} \right\|^2 + \lambda \sum_{j=1}^{s} \|\vec{\beta}_j\|^2 + \sum_{j=1}^{s} \lambda_j' \|\vec{\beta}_j\|_1 \right]$$