PCA and MDS
A Geometric View

Yuan Yao

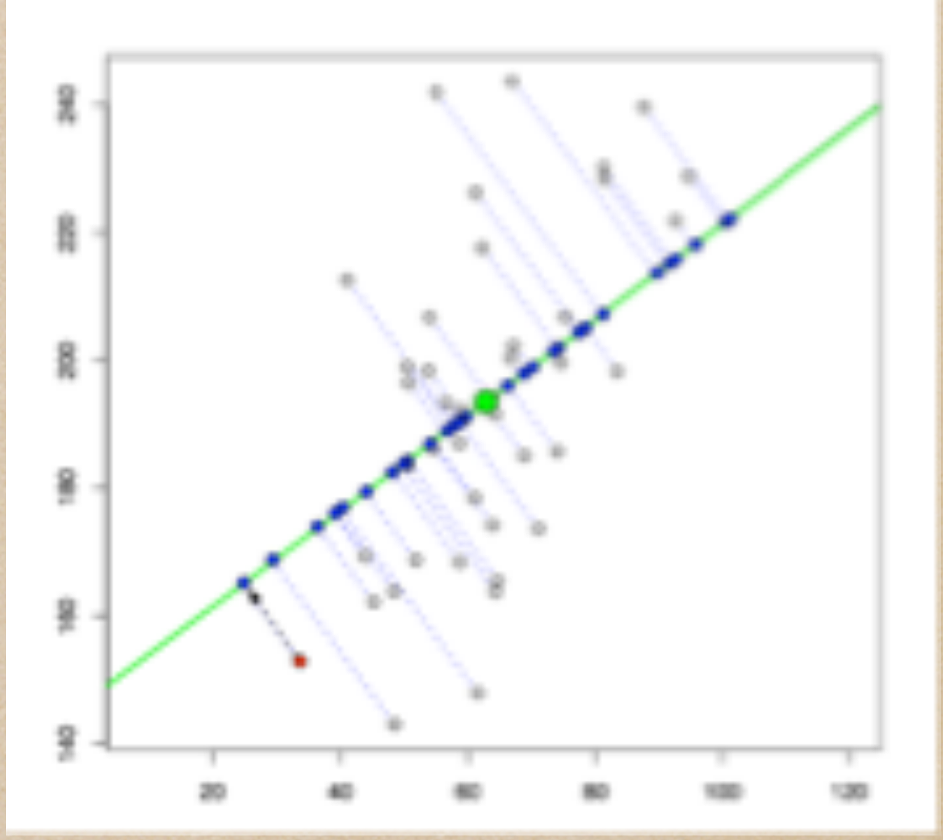
Geometric Embedding

- A Fundamental Problem in Data Representation
- Unstructured data -> Euclidean Space
    - PCA: high dim -> low dim affine space
    - MDS: metric -> Euclidean space
- a.k.a. `feature' learning (e.g. deep learning)
- speech, text, image, video...

Principal Component Analysis (PCA)

Let $x_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, be n samples in $\mathbb{R}^p$.

- Can you find a low dimensional affine representation?



Principal Component Analysis (PCA)

Best k-affine space approximation of data:

Let $\boldsymbol{X} = [\boldsymbol{X}_1 | \boldsymbol{X}_2 | \cdots | \boldsymbol{X}_n] \in \mathbb{R}^{p \times n}$.

(2) $$\min_{\beta, \mu, U} I := \sum_{i=1}^{n} \|X_i - (\mu + U\beta_i)\|^2$$

where $U \in \mathbb{R}^{p \times k}$, $U^T U = I_p$, and $\sum_{i=1}^{n} \beta_i = 0$ (nonzero sum of β_i can be repre-

Plug in the expression of $\hat{\mu}_n$ and β_i

$$\begin{aligned} I &= \sum_{i=1}^{n} \|X_i - \hat{\mu}_n - UU^T(X_i - \hat{\mu}_n)\|^2 \\ &= \sum_{i=1}^{n} \|X_i - \hat{\mu}_n - P_k(X_i - \hat{\mu}_n)\|^2 \\ &= \sum_{i=1}^{n} \|Y_i - P_k(y_i)\|^2, \quad Y_i := X_i - \hat{\mu}_n \end{aligned}$$

where $P_k = UU^T$ is a projection operator satisfying the idempotent property $P_k^2 = P_k$.

Denote $Y = [Y_1|Y_2|\cdots|Y_n] \in \mathbb{R}^{p\times n}$, whence the original problem turns into

$$\begin{aligned}
\min_U \sum_{i=1}^n \|Y_i - P_k(Y_i)\|^2 &= \min \operatorname{trace}[(Y - P_kY)^T(Y - P_kY)] \\
&= \min \operatorname{trace}[Y^T(I - P_k)(I - P_k)Y] \\
&= \min \operatorname{trace}[YY^T(I - P_k)^2] \\
&= \min \operatorname{trace}[YY^T(I - P_k)] \\
&= \min[\operatorname{trace}(YY^T) - \operatorname{trace}(YY^TUU^T)] \\
&= \min[\operatorname{trace}(YY^T) - \operatorname{trace}(U^TYY^TU)].
\end{aligned}$$

Above we use cyclic property of trace and idempotent property of projection.

Since Y does not depend on U, the problem above is equivalent to

$$(3) \qquad \max_{UU^T=I_k} Var(U^TY) = \max_{UU^T=I_k} \frac{1}{n}\operatorname{trace}(U^TYY^TU) = \max_{UU^T=I_k} \operatorname{trace}(U^T\hat{\Sigma}_nU)$$

where $\hat{\Sigma}_n = \frac{1}{n}YY^T = \frac{1}{n}(X - \hat{\mu}_n\mathbf{1}^T)(X - \hat{\mu}_n\mathbf{1}^T)^T$ is the sample variance. Assume

$$\frac{\partial I}{\partial \mu} = -2\sum_{i=1}^{n}(X_i - \mu - U\beta_i) = 0 \Rightarrow \hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\frac{\partial I}{\partial \beta_i} = (x_i - \mu - U\beta_i)^T U = 0 \Rightarrow \beta_i = U^T(X_i - \mu)$$

Principal Component Analysis Summary

- PCA is given by the top k eigenvector of covariance matrix

$$\hat{\Sigma}_n = \frac{1}{n-1}\tilde{X} \cdot \tilde{X}^T$$

$$\tilde{X} = XH = X - \frac{1}{n}X \cdot \mathbf{1}\mathbf{1}^T = \tilde{U}\tilde{S}\tilde{V}^T, \quad H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T,\ \mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$$

(left) singular vectors here gives PCA

How much variances in data explained by PCA?

- total variance:

$$\mathrm{trace}(\hat{\Sigma}_n) = \sum_{i=1}^{p} \hat{\lambda}_i;$$

- percentage of variance explained by top-k principal components:

$$\sum_{i=1}^{k} \hat{\lambda}_i / \mathrm{trace}(\hat{\Sigma}_n);$$

- generalized variance as total volume:

$$\det(\hat{\Sigma}_n) = \prod_{i=1}^{p} \hat{\lambda}_i.$$

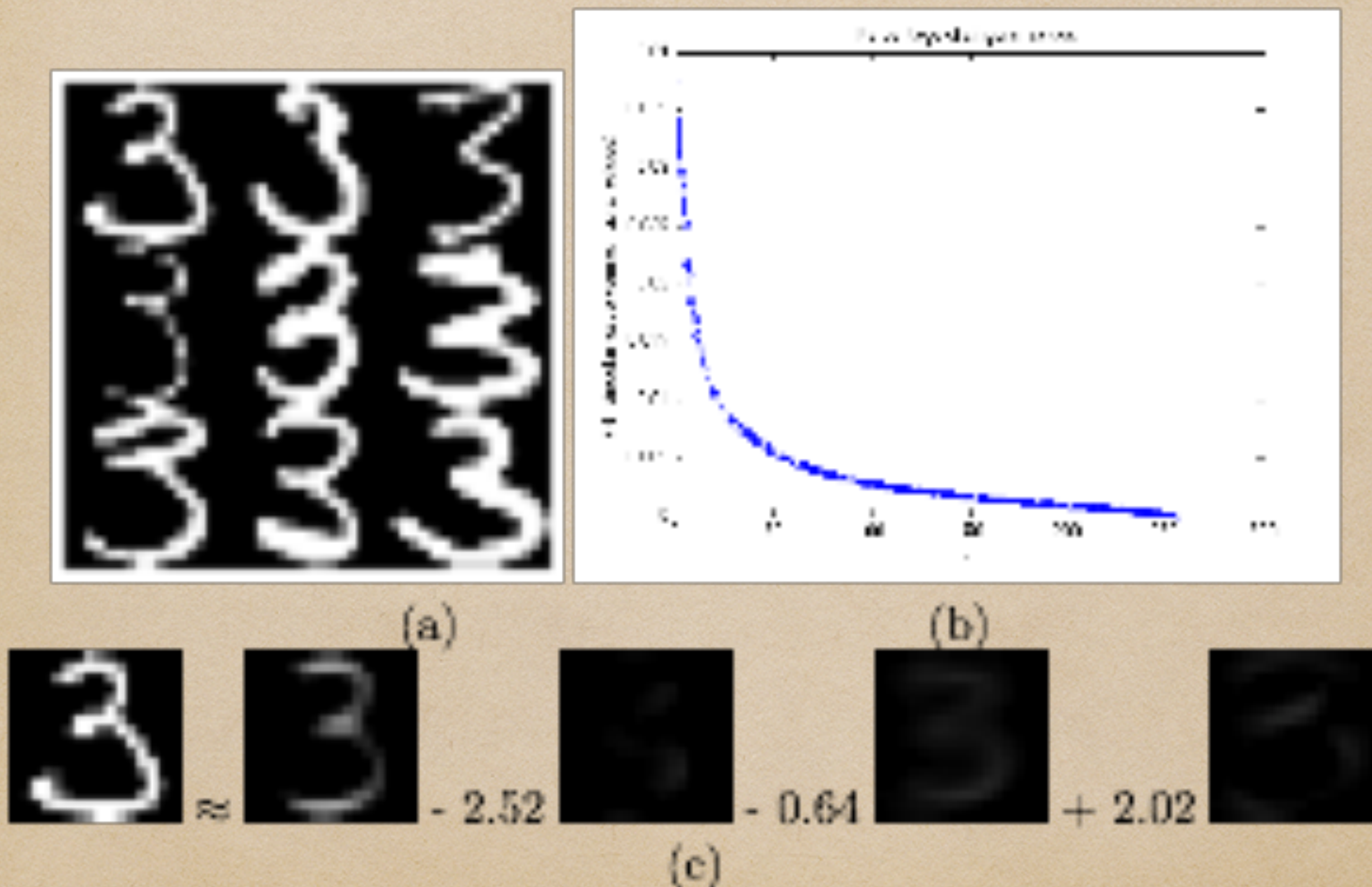
Example:

- Choose k such that

$$\sum_{i=1}^{k} \hat{\lambda}_i / \mathrm{trace}(\hat{\Sigma}_n) > 0.95.$$

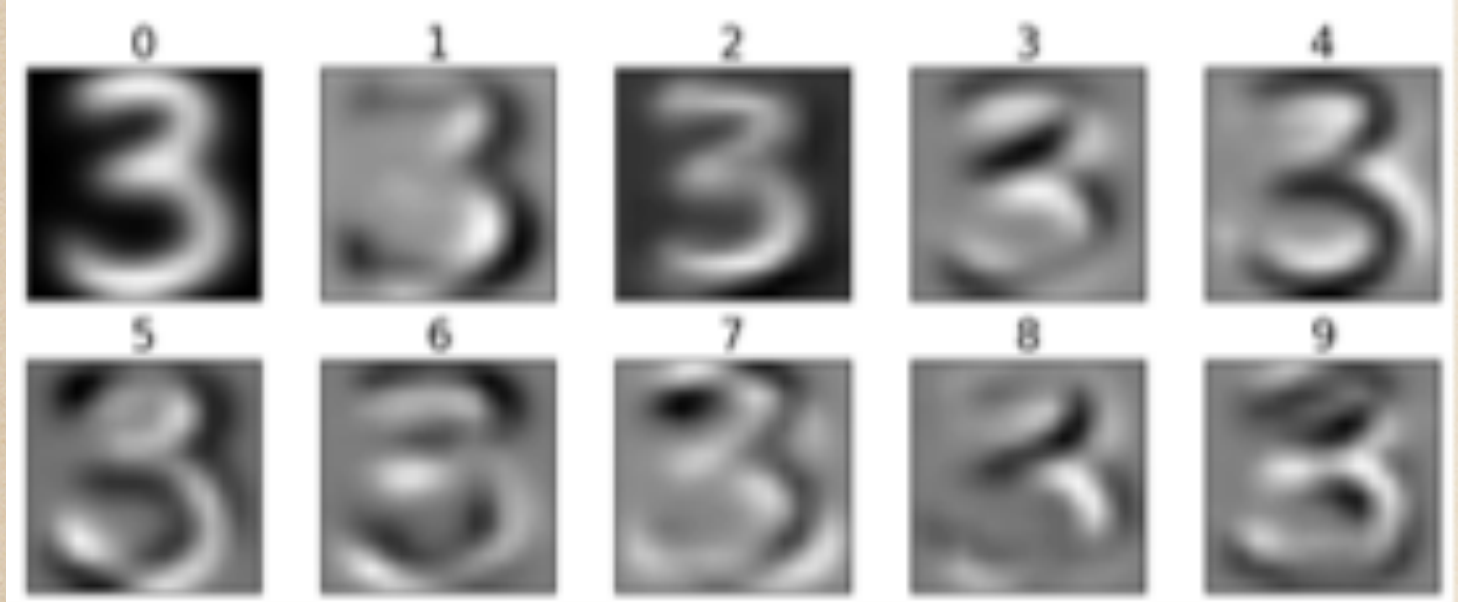
explains 95% total variations in data

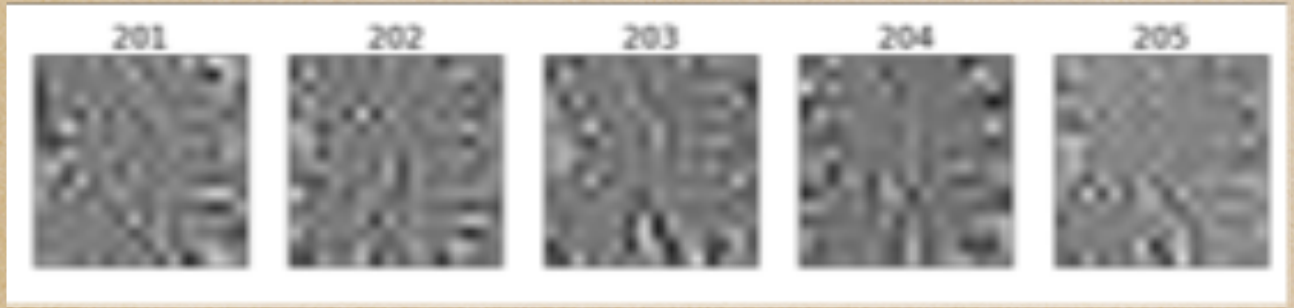
Example of PCA



Mean and PCs

Mean and top 9 PCs that explains >98% total variations





PC 201-205 are less informative

Random Permutation Test

- a.k.a. Horn's Parallel Analysis
- randomly permute samples for decorrelation
- compute eigenvalues of random matrices

$$X = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,n} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p,1} & X_{p,2} & \cdots & X_{p,n} \end{bmatrix} \longrightarrow X^1 = \begin{bmatrix} X_{1,\pi_1(1)} & X_{1,\pi_1(2)} & \cdots & X_{1,\pi_1(n)} \\ X_{2,\pi_2(1)} & X_{2,\pi_2(2)} & \cdots & X_{2,\pi_2(n)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{p,\pi_p(1)} & X_{p,\pi_p(2)} & \cdots & X_{p,\pi_p(n)} \end{bmatrix}$$

$$\{\hat{\lambda}_i^1\}_{i=1,\ldots,p}$$

Horn's Parallel Analysis

- Repeat this for R times, obtain a R-by-p eigenvalue matrix

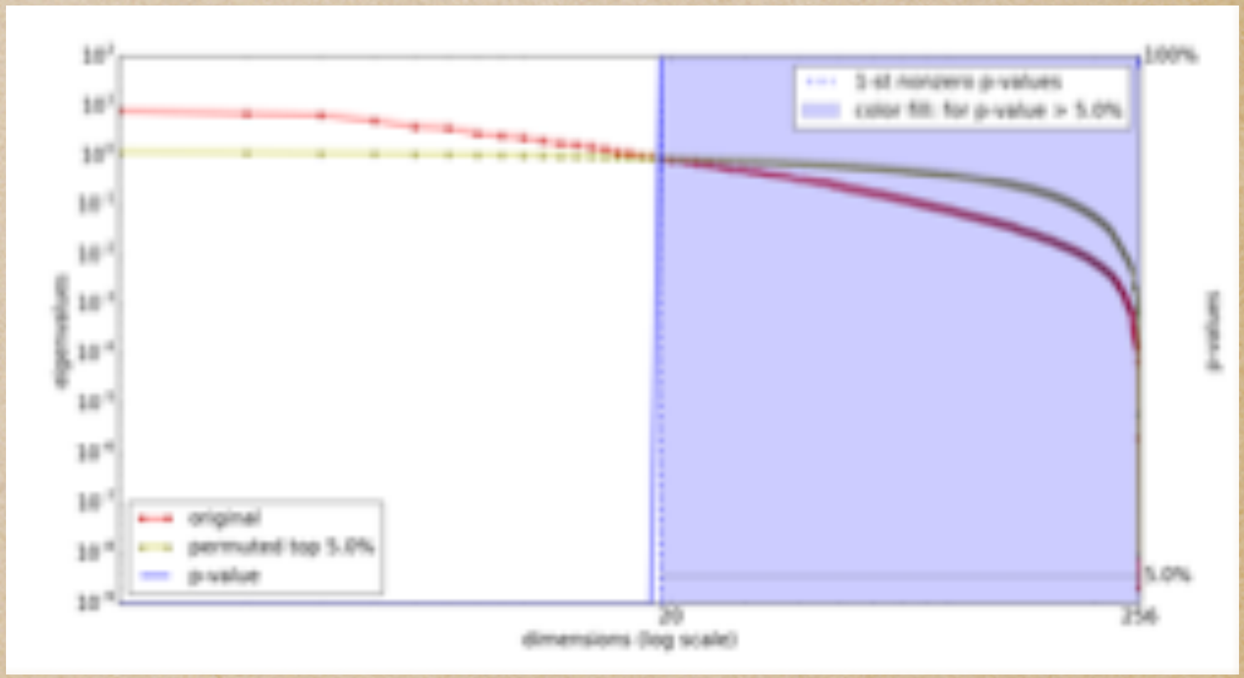
$$\begin{bmatrix} \tilde{\lambda}_1^1 & \tilde{\lambda}_2^1 & \cdots & \tilde{\lambda}_p^1 \\ \tilde{\lambda}_1^2 & \tilde{\lambda}_2^2 & \cdots & \tilde{\lambda}_p^2 \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{\lambda}_1^R & \tilde{\lambda}_2^R & \cdots & \tilde{\lambda}_p^R \end{bmatrix}$$

- Define the p-value for the i-th eigenvalue, and only keep eigenvalues whose pval is smaller than a threshold, e.g.

$$\text{pval}_i = \frac{1}{R}\#\{\tilde{\lambda}_i^r > \hat{\lambda}_i\} \qquad \text{pval}_i < 0.05$$

Example

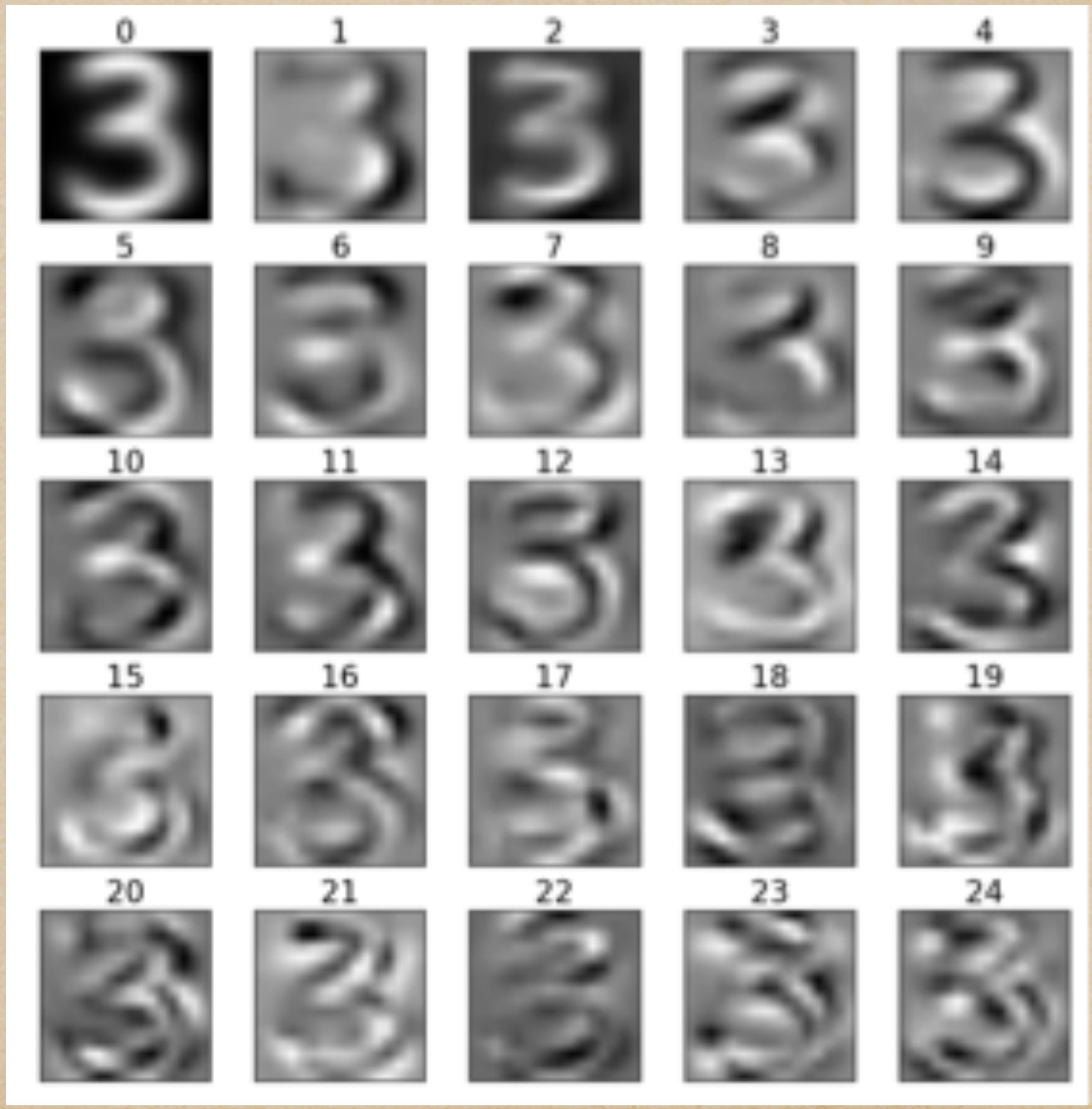
- Fig 1. Examples of randomly permuted data



- Fig 2. Results of parallel analysis on digit 3, with R=100. There are 19 PCs whose pval's <5%

Example

- Top 24 PCs and top 19 is suggested by parallel analysis at 5% level, which might be conservative as the images lie near a submanifold in the space



Summary: PCA=SVD

(SVD) of $X = [x_1, \dots, x_n]^T \in \mathbb{R}^{n \times p}$ in the following sense,

$$Y = X - \frac{1}{n}\mathbf{1}\mathbf{1}^T X = \tilde{U}\tilde{S}\tilde{V}^T, \quad \mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}^n$$

[2] In statistics, data matrix is often written as samples row-wise and variables column-wise, i.e. n-by-p matrix. So be careful on your way of writing the data matrix!

- top k right singular vectors give PCA (Covariance spectrum)
- top k left singular vectors? It gives MDS (Kernel spectrum)