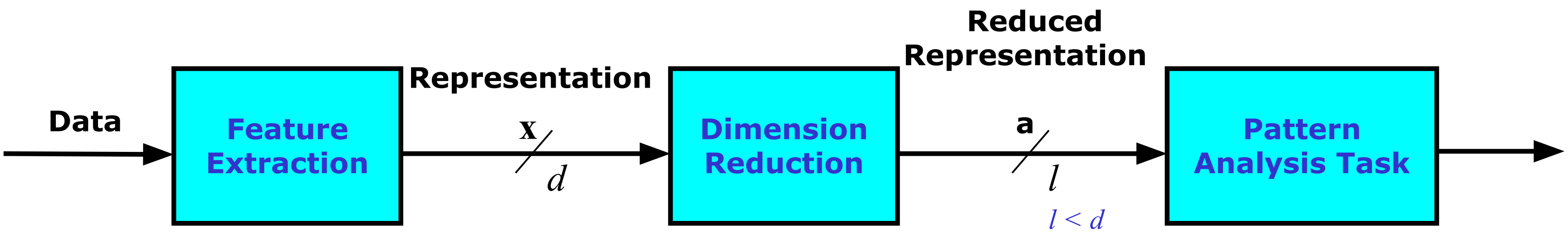
Dimensionality Reduction
Principal Component Analysis (PCA)

Dimensionality Reduction

- Data encoding or transformations are applied so as to obtain a reduced or compressed representation of the original data



- If the original data can be reconstructed from compressed data without any loss of information, the data reduction is called lossless
- If only an approximation of the original data can be reconstructed from compressed data, then the data reduction is called lossy
- One of the popular and effective methods of lossy dimensionality reduction is principal component analysis (PCA)

Principal Component Analysis (PCA)

- Suppose data to be reduced consist of N tuples (or data vectors) described by d-attributes (d-dimensions)

$$\mathrm{D} = \{\mathbf{x}_n\}_{n=1}^{N}, \mathbf{x}_n \in \mathrm{R}^d$$

$$\mathbf{x}_n = [x_{n1} \;\; x_{n2} \;\; \ldots \;\; x_{nd}]^{\mathsf{T}}$$

- Let $\mathbf{q}_i$, where $i = 1, 2, \ldots, d$ be the d orthonormal vectors in the d –dimensional space, $\mathbf{q}_i \in \mathrm{R}^d$
  - These are unit vectors that each point in a direction perpendicular to the others

$$\mathbf{q}_i^{\mathsf{T}} \mathbf{q}_j = 0 \quad \forall i \neq j$$

$$\mathbf{q}_i^{\mathsf{T}} \mathbf{q}_i = 1$$

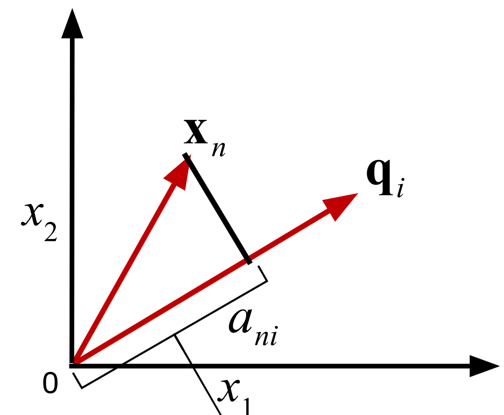
- These orthonormal vectors are also called as direction of projection

Principal Component Analysis (PCA)

- PCA searches for l orthonormal vectors that can best be used to represent the data, where $l < d$
- The original data (each of the tuples (data vectors), $\mathbf{x}_n$) is then projected onto each of the l orthonormal vectors get the principal components

$$a_{ni} = \mathbf{q}_i^{\mathsf{T}} \mathbf{x}_n \quad \forall i = 1, 2, \ldots, l$$

  – a_{ni} is an i^{th} principal component of $\mathbf{x}_n$



- This transform each of the d – dimensional vectors (i.e. tuples) to l – dimensional vectors

$$\mathbf{x}_n = \begin{bmatrix} x_{n1} \\ x_{n2} \\ \cdots \\ x_{nd} \end{bmatrix} \rightarrow \mathbf{a}_n = \begin{bmatrix} a_{n1} \\ a_{n2} \\ \cdots \\ a_{nl} \end{bmatrix}$$

- Task:
  – How to obtain the orthonormal vectors?
  – Which l orthonormal vectors to choose from d orthonormal vectors?

PCA: Basic Procedure

- Given: Data with N samples, D $= \{\mathbf{x}_n\}_{n=1}^{N}, \mathbf{x}_n \in \mathrm{R}^d$

1. Remove mean for each attribute (dimension) in data samples (tuples)
2. Then construct a data matrix $\mathbf{X}$ using the mean subtracted samples, $\mathbf{X} \in \mathrm{R}^{N \times d}$
   – Each row of the matrix $\mathbf{X}$ corresponds to 1 sample (tuple)
3. Compute a correlation matrix $\mathbf{C}=\mathbf{X}^T\mathbf{X}$
   - This correlation matrix is equivalent to covariance matrix from original data matrix:

$$\mathbf{C}\mathbf{q}_i = \lambda_i \mathbf{q}_i \quad \forall i = 1, 2, ..., d$$

4. Perform the eigen analysis of covariance matrix $\mathbf{C}$
   1. As covariance matrix is symmetric matrix,
      1. Each eigenvalues λ_i are distinct and non-negative
      - Eigenvectors $\mathbf{q}_i$ corresponding to each eigenvalues are orthonormal vectors
      1. Eigenvalues indicate the strength of eigenvectors or variance of projected data in the direction of eigenvector

PCA for Dimension Reduction

- In general, we are interested in representing the data using fewer dimensions such that the data has high variance along these dimensions

5. Rank order the eigenvalues (λ_i's) (descending order of λ_i's) such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$

6. Consider the l ($l << d$) eigenvectors corresponding to l significant eigenvalues

7. Project the $\mathbf{x}_n$ onto each of the l directions (eigenvectors) to get reduced dimensional representation in terms of principal components

$$a_{ni} = \mathbf{q}_i^{\mathsf{T}} \mathbf{x}_n \quad \forall i = 1, 2, \ldots, l$$

a_{ni} is an i^{th} principal component of $\mathbf{x}_n$

PCA for Dimension Reduction

8. Thus, each training example $\mathbf{x}_n$ is transformed to a new reduced dimensional representation $\mathbf{a}_n$ by projecting on to l-orthonormal basis

$$\mathbf{x}_n = \begin{bmatrix} x_{n1} \\ x_{n2} \\ \cdots \\ x_{nd} \end{bmatrix} \rightarrow \mathbf{a}_n = \begin{bmatrix} a_{n1} \\ a_{n2} \\ \cdots \\ a_{nl} \end{bmatrix}$$

- Observations:
  - The new reduced representation $\mathbf{a}_n$ is uncorrelated
  - The eigenvalue λ_i correspond to the variance of projected data (reduced representation)
- Note:
  - The number l is chosen experimentally (empirically) by observing the values of eigenvalue
  - If the data is projected onto all the eigenvectors, $\mathbf{x}_n$ is transformed to a new representation $\mathbf{a}_n$ with d-dimension

    The new representation is uncorrelated

Illustration: PCA

Temperature	Humidity	Pressure	Rain	Moisture
25.47	82.19	1036.35	6.75	0.00
26.19	83.15	1037.60	1761.75	5.69
25.17	85.34	1037.89	652.50	6.85
24.30	87.69	1036.86	963.00	6.04
24.07	87.65	1027.83	254.25	31.24
21.21	95.95	1006.92	339.75	100.00
23.49	96.17	1006.57	38.25	93.20
21.79	98.59	1009.42	29.25	5.77
25.09	88.33	991.65	4.50	4.29
25.39	90.43	1009.66	112.50	3.62
23.89	94.54	1009.27	735.75	3.76
22.51	99.00	1009.80	607.50	4.03
22.90	98.00	1009.90	717.75	3.83
21.72	99.00	996.29	513.00	3.04
23.18	98.97	800.00	195.75	3.00
21.24	99.00	1009.21	474.75	3.05
21.63	99.00	1008.89	409.50	3.00
20.91	99.00	1008.89	1161.00	3.20
23.67	97.80	1009.38	0.00	2.04
24.53	92.90	1008.66	0.00	1.80

- Atmospheric Data:
  - N = Number tiples (data vectors) = 20
  - d = Number of attributes (dimension) = 5
- Mean of each dimension:

23.42	93.64	1003.55	448.88	14.4

Illustration: PCA

Temperature	Humidity	Pressure	Rain	Moisture
2.05	-11.45	32.80	-442.13	-14.37
2.77	-10.49	34.05	1312.88	-8.68
1.75	-8.30	34.34	203.63	-7.52
0.88	-5.95	33.31	514.13	-8.33
0.65	-5.99	24.28	-194.63	16.87
-2.21	2.32	3.37	-109.13	85.63
0.07	2.54	3.02	-410.63	78.83
-1.63	4.96	5.87	-419.63	-8.60
1.67	-5.31	-11.90	-444.38	-10.08
1.97	-3.21	6.11	-336.38	-10.75
0.47	0.91	5.72	286.88	-10.61
-0.91	5.36	6.25	158.63	-10.34
-0.52	4.36	6.35	268.88	-10.54
-1.70	5.36	-7.26	64.13	-11.33
-0.24	5.33	-203.55	-253.13	-11.37
-2.18	5.36	5.66	25.88	-11.32
-1.79	5.36	5.34	-39.38	-11.37
-2.51	5.36	5.34	712.13	-11.17
0.25	4.16	5.83	-448.88	-12.33
1.11	-0.73	5.11	-448.88	-12.57

- Step1: Subtract mean from each attribute

Illustration: PCA

- Step2: Multiply the mean subtracted data matrix with its transpose (i.e., covariance matrix of original data matrix)

2.64	-8.21	14.13	16.98	-9.66
-8.21	35.05	-117.00	-459.95	13.34
14.13	-117.00	2478.61	5420.36	80.09
16.98	-459.95	5420.36	215276.95	-2427.97
-9.66	13.34	80.09	-2427.97	832.14

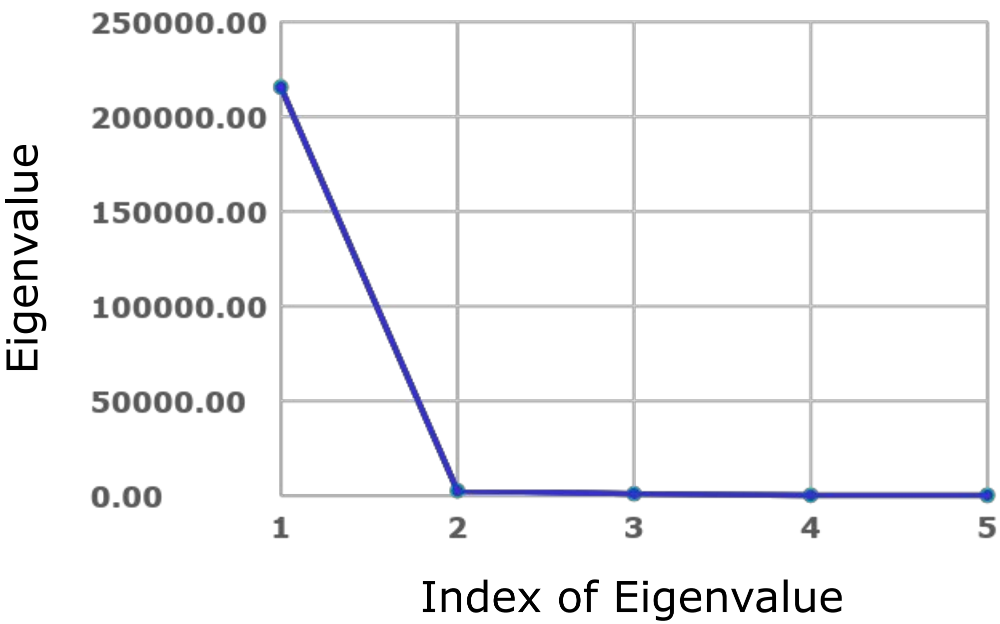
Illustration: PCA

Eigen Values

215443.33	2358.36	792.30	30.88	0.52

Eigen Vectors

-0.0001	0.0056	-0.0137	0.2498	0.9682
0.0021	-0.0448	0.0232	-0.9669	0.2501
-0.0254	0.9946	-0.0892	-0.0469	0.0051
-0.9996	-0.0244	0.0136	-0.0007	0.0004
0.0113	0.0906	0.9956	0.0218	0.0080



- Step3: Perform Eigen analysis on correlation matrix
  - Get eigenvalues and eigenvectors
- Step4: Sort the eigenvalues in descending order
- Step5: Arrange the eigenvectors in the descending order of their corresponding eigenvalues
- Step6: Consider the two leading (significant) eigenvalues and their corresponding eigenvectors

Illustration: PCA

a_1	a_2
440.93	42.62
-1313.35	1.55
-204.52	28.89
-514.88	20.11
194.11	30.69
109.97	13.65
411.28	20.04
419.23	15.05
444.38	-1.66
335.96	13.46
-287.03	-2.31
-158.83	1.16
-269.04	-1.40
-64.03	-10.06
258.09	-197.54
-26.13	3.72
39.11	4.99
-712.10	-13.32
448.42	15.44
448.43	14.93

- Step7: Project the mean subtracted data matrix onto the selected two eigenvectors corresponding to leading eigenvalues

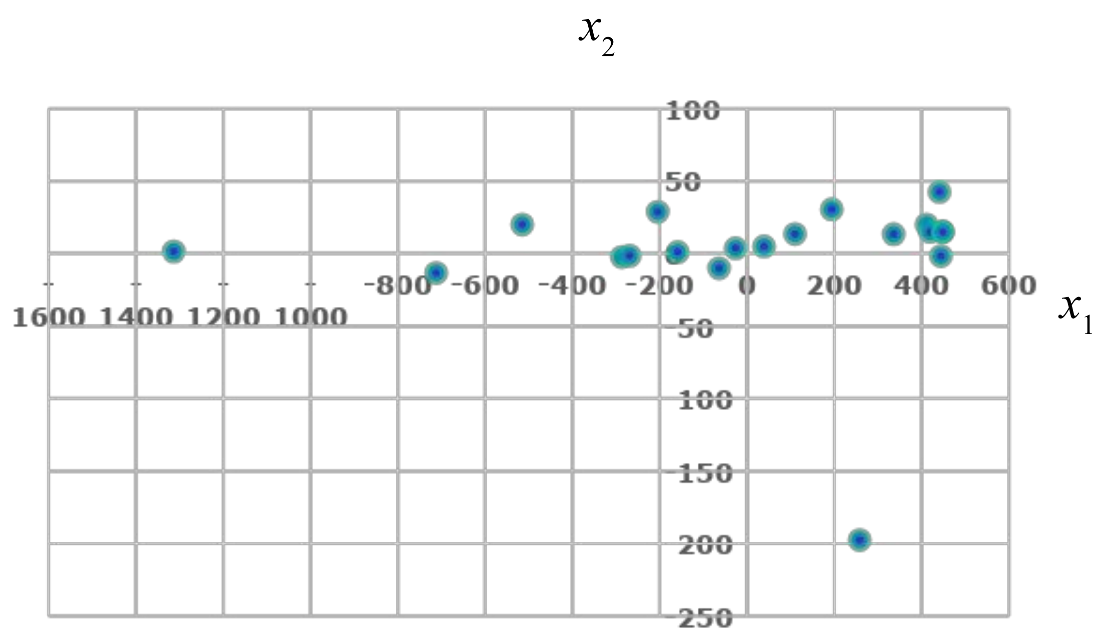


Illustration: PCA

a_1	a_2
440.93	42.62
-1313.35	1.55
-204.52	28.89
-514.88	20.11
194.11	30.69
109.97	13.65
411.28	20.04
419.23	15.05
444.38	-1.66
335.96	13.46
-287.03	-2.31
-158.83	1.16
-269.04	-1.40
-64.03	-10.06
258.09	-197.54
-26.13	3.72
39.11	4.99
-712.10	-13.32
448.42	15.44
448.43	14.93

- Step7: Project the mean subtracted data matrix onto the selected two eigenvectors corresponding to leading eigenvalues

- Covariance matrix of 2-dimensional representation obtained using PCA:

215443.33	0.00
0.00	2358.36

- The reduced representation is uncorrelated

Illustration: PCA – Reconstruction of Data

x_1	x_2	x_3	x_4	x_5	$Error$
0.20	-0.96	31.17	-441.80	8.84	25.59
0.12	-2.89	34.96	1312.80	-14.70	10.09
0.18	-1.73	33.93	203.74	0.30	10.34
0.15	-2.01	33.09	514.19	-4.00	5.91
0.16	-0.96	25.59	-194.78	4.97	12.99
0.07	-0.37	10.78	-110.26	2.48	83.56
0.08	-0.01	9.47	-411.61	6.46	72.70
0.05	0.23	4.31	-419.43	6.10	15.61
-0.05	1.03	-12.96	-444.17	4.87	16.37
0.05	0.12	4.84	-336.16	5.01	16.28
0.01	-0.51	5.01	286.97	-3.45	7.34
0.02	-0.39	5.20	158.74	-1.69	10.49
0.01	-0.52	5.45	268.97	-3.17	8.90
-0.05	0.31	-8.38	64.25	-1.63	11.11
-1.12	9.40	-203.04	-253.17	-14.97	5.52
0.02	-0.22	4.36	26.02	0.04	12.92
0.02	-0.14	3.97	-39.21	0.89	13.64
-0.02	-0.93	4.87	712.14	-9.25	7.06
0.05	0.27	3.95	-448.62	6.47	19.30
0.05	0.30	3.44	-448.62	6.42	19.12

- An approximation of mean subtracted data, $\mathbf{x}_n$, is obtained as linear combination of the direction of projection (strongest eigenvectors), $\mathbf{q}_i$, and the principal components, a_{ni}

$$\hat{\mathbf{X}}_n = \sum_{i=1}^{l} a_{ni}\mathbf{q}_i$$

- Error in reconstruction: The Euclidean distance between the original and approximated tuples

Illustration: PCA - Extended

a_1	a_2	a_3	a_4	a_5
440.93	42.62	-23.53	10.02	-1.01
-1313.35	1.55	5.86	8.18	0.72
-204.52	28.89	-8.00	6.55	-0.18
-514.88	20.11	-4.44	3.89	-0.32
194.11	30.69	11.84	5.31	-0.69
109.97	13.65	83.55	-1.01	-0.91
411.28	20.04	72.69	-0.59	1.17
419.23	15.05	-14.65	-5.38	-0.55
444.38	-1.66	-15.16	6.18	-0.03
335.96	13.46	-15.92	3.29	0.92
-287.03	-2.31	-7.16	-1.44	0.75
-158.83	1.16	-8.56	-6.04	0.48
-269.04	-1.40	-7.30	-5.05	0.66
-64.03	-10.06	-9.61	-5.56	-0.40
258.09	-197.54	3.52	4.25	-0.13
-26.13	3.72	-11.27	-6.26	-0.81
39.11	4.99	-12.18	-6.11	-0.47
-712.10	-13.32	-1.78	-6.78	-0.85
448.42	15.44	-18.80	-4.21	1.03
448.43	14.93	-19.10	0.77	0.63

- Step6: Consider the all the eigenvalues and their corresponding eigenvectors
- Step7: Project the mean subtracted data matrix onto all the eigenvectors
- The resultant 5-dimensional representation is a new transformed representation
- Covariance matrix of new representation obtained using PCA:

215443.33	0.00	0.00	0.00	0.00
0.00	2358.36	0.00	0.00	0.00
0.00	0.00	792.30	0.00	0.00
0.00	0.00	0.00	30.88	0.00
0.00	0.00	0.00	0.00	0.52

- The encoded representation is uncorrelated

Illustration: PCA – Reconstruction of Data

x_1	x_2	x_3	x_4	x_5	*Error*
2.05	-11.45	32.80	-442.13	-14.37	1.79E-14
2.77	-10.49	34.05	1312.88	-8.68	3.83E-14
1.75	-8.29	34.34	203.63	-7.52	2.18E-14
0.88	-5.95	33.31	514.13	-8.33	1.15E-13
0.65	-5.99	24.28	-194.63	16.87	3.43E-14
-2.21	2.31	3.37	-109.13	85.63	1.55E-14
0.07	2.54	3.02	-410.63	78.83	5.73E-14
-1.62	4.96	5.86	-419.63	-8.60	5.87E-14
1.68	-5.31	-11.90	-444.38	-10.08	5.83E-14
1.98	-3.20	6.11	-336.38	-10.76	1.19E-14
0.47	0.90	5.72	286.88	-10.61	5.73E-14
-0.91	5.37	6.24	158.63	-10.34	4.07E-15
-0.51	4.37	6.34	268.88	-10.54	5.73E-14
-1.69	5.37	-7.26	64.13	-11.33	1.52E-14
-0.24	5.34	-203.55	-253.13	-11.37	1.17E-13
-2.18	5.37	5.65	25.88	-11.32	2.66E-15
-1.79	5.37	5.34	-39.38	-11.37	3.23E-15
-2.51	5.37	5.34	712.13	-11.18	1.15E-13
0.25	4.17	5.83	-448.88	-12.34	9.06E-15
1.11	-0.73	5.11	-448.88	-12.57	5.82E-14

- An approximation of mean subtracted data, $\mathbf{x}_n$, is obtained as linear combination of the direction of projection (strongest eigenvectors), $\mathbf{q}_i$, and the principal components, a_{ni}

$$\hat{\mathbf{x}}_n = \sum_{i=1}^{d} a_{ni}\mathbf{q}_i$$

- Error in reconstruction: The Euclidean distance between the original and approximated tuples

Illustration: PCA – *Projecting Original Data*

a_1	a_2
-32.94	1027.02
-1787.22	985.95
-678.39	1013.28
-988.75	1004.50
-279.76	1015.09
-363.90	998.05
-62.58	1004.44
-54.64	999.45
-29.49	982.74
-137.91	997.85
-760.89	982.09
-632.69	985.56
-742.91	983.00
-537.90	974.34
-215.78	786.85
-499.99	988.11
-434.76	989.39
-1185.97	971.08
-25.45	999.84
-25.44	999.32

- Step6: Consider the two leading (significant) eigenvalues and their corresponding eigenvectors
- Step7: Project the original data matrix *(not mean subtracted)* onto the selected two eigenvectors corresponding to leading eigenvalues
- Covariance matrix of 2-dimensional representation obtained using PCA:

215443.33	0.00
0.00	2358.36

- The reduced representation is uncorrelated

Illustration: PCA – *Projecting Original Data -* Reconstruction of Data

x_1	x_2	x_3	x_4	x_5	*Error*
5.75	-46.08	1022.31	7.87	92.68	160.09
5.70	-47.92	1026.02	1762.45	69.13	147.51
5.74	-46.82	1025.04	653.39	84.14	154.87
5.72	-47.08	1024.19	963.84	79.83	155.29
5.71	-46.06	1016.71	254.88	88.81	147.15
5.63	-45.48	1001.90	339.40	86.31	143.03
5.63	-45.13	1000.61	38.05	90.30	142.58
5.60	-44.89	995.44	30.23	89.93	167.71
5.51	-44.09	978.18	5.50	88.70	158.83
5.60	-44.99	995.96	113.51	88.85	161.81
5.58	-45.60	996.11	736.62	80.38	161.30
5.58	-45.48	996.31	608.39	82.14	165.67
5.58	-45.60	996.56	718.63	80.66	164.32
5.51	-44.78	982.74	513.91	82.20	165.49
4.43	-35.70	788.08	196.49	68.85	151.55
5.58	-45.32	995.47	475.68	83.87	166.72
5.58	-45.24	995.09	410.44	84.73	167.13
5.56	-45.99	995.96	1161.80	74.58	162.85
5.60	-44.85	995.09	1.04	90.30	169.32
5.60	-44.82	994.57	1.05	90.25	165.37

- An approximation of original data, $\mathbf{x}_n$, is obtained as linear combination of the direction of projection (strongest eigenvectors), $\mathbf{q}_i$, and the principal components, a_{ni}

$$\hat{\mathbf{X}}_n = \sum_{i=1}^{l} a_{ni}\mathbf{q}_i$$

- Error in reconstruction: The Euclidean distance between the original and approximated tuples

Illustration: PCA – Extended - *Projecting Original Data*

a_1	a_2	a_3	a_4	a_5
-32.94	1027.02	-90.78	-121.73	50.52
-1787.22	985.95	-61.39	-123.57	52.25
-678.39	1013.28	-75.26	-125.20	51.35
-988.75	1004.50	-71.69	-127.86	51.21
-279.76	1015.09	-55.42	-126.44	50.84
-363.90	998.05	16.30	-132.76	50.62
-62.58	1004.44	5.44	-132.34	52.70
-54.64	999.45	-81.90	-137.13	50.98
-29.49	982.74	-82.41	-125.57	51.49
-137.91	997.85	-83.17	-128.46	52.44
-760.89	982.09	-74.41	-133.19	52.28
-632.69	985.56	-75.82	-137.79	52.01
-742.91	983.00	-74.55	-136.80	52.18
-537.90	974.34	-76.87	-137.31	51.13
-215.78	786.85	-63.73	-127.50	51.40
-499.99	988.11	-78.52	-138.01	50.71
-434.76	989.39	-79.44	-137.86	51.06
-1185.97	971.08	-69.03	-138.53	50.67
-25.45	999.84	-86.05	-135.96	52.55
-25.44	999.32	-86.35	-130.98	52.16

- Step6: Consider the all the eigenvalues and their corresponding eigenvectors
- Step7: Project the original data matrix *(not mean subtracted)* onto all the eigenvectors
- The resultant 5-dimensional representation is a new transformed representation
- Covariance matrix of new representation obtained using PCA:

215443.33	0.00	0.00	0.00	0.00
0.00	2358.36	0.00	0.00	0.00
0.00	0.00	792.30	0.00	0.00
0.00	0.00	0.00	30.88	0.00
0.00	0.00	0.00	0.00	0.52

- The encoded representation is uncorrelated

Illustration: PCA – *Projecting Original Data -* Reconstruction of Data

x_1	x_2	x_3	x_4	x_5	*Error*
25.50	82.15	1036.38	6.74	0.05	0.08
26.26	83.20	1037.56	1761.72	5.74	0.11
25.22	85.33	1037.89	652.48	6.89	0.07
24.35	87.69	1036.84	962.98	6.08	0.07
24.11	87.62	1027.85	254.24	31.28	0.07
21.25	95.93	1006.93	339.74	100.05	0.07
23.52	96.14	1006.60	38.23	93.25	0.07
21.83	98.55	1009.44	29.23	5.81	0.07
25.12	88.29	991.69	4.49	4.33	0.08
25.42	90.40	1009.68	112.49	3.66	0.06
23.94	94.53	1009.26	735.72	3.81	0.08
22.56	99.00	1009.80	607.48	4.07	0.07
22.95	97.99	1009.89	717.73	3.88	0.07
21.77	98.99	996.30	512.98	3.08	0.07
23.22	98.95	800.01	195.74	3.03	0.06
21.28	98.99	1009.21	474.73	3.10	0.07
21.67	98.99	1008.90	409.48	3.04	0.06
20.96	99.02	1008.87	1160.98	3.24	0.07
23.70	97.76	1009.41	-0.01	2.08	0.07
24.56	92.86	1008.68	-0.02	1.84	0.07

- An approximation of original data, $\mathbf{x}_n$, is obtained as linear combination of the direction of projection (strongest eigenvectors), $\mathbf{q}_i$, and the principal components, a_{ni}

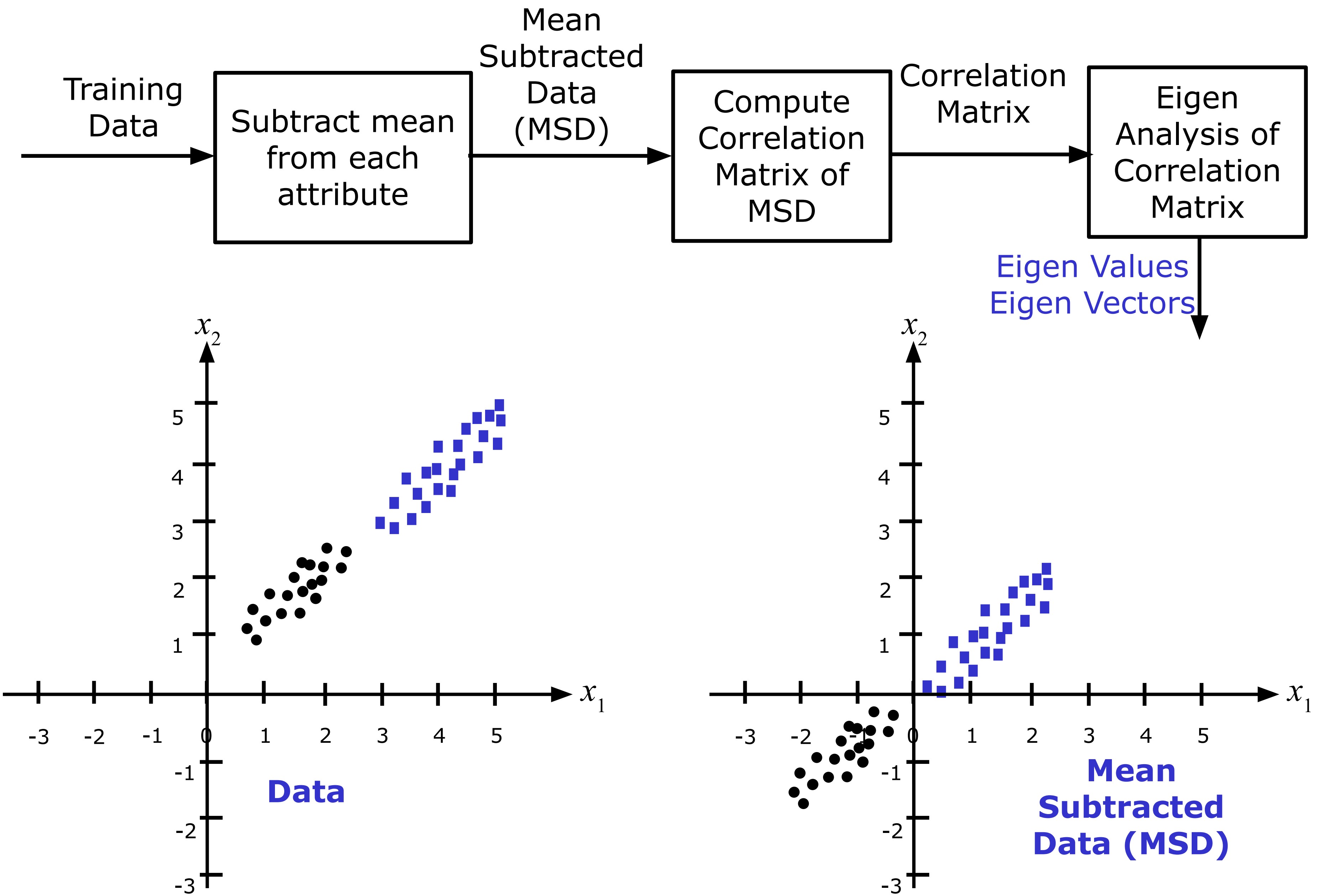
$$\hat{\mathbf{X}}_n = \sum_{i=1}^{l} a_{ni}\mathbf{q}_i$$

- Error in reconstruction: The Euclidean distance between the original and approximated tuples

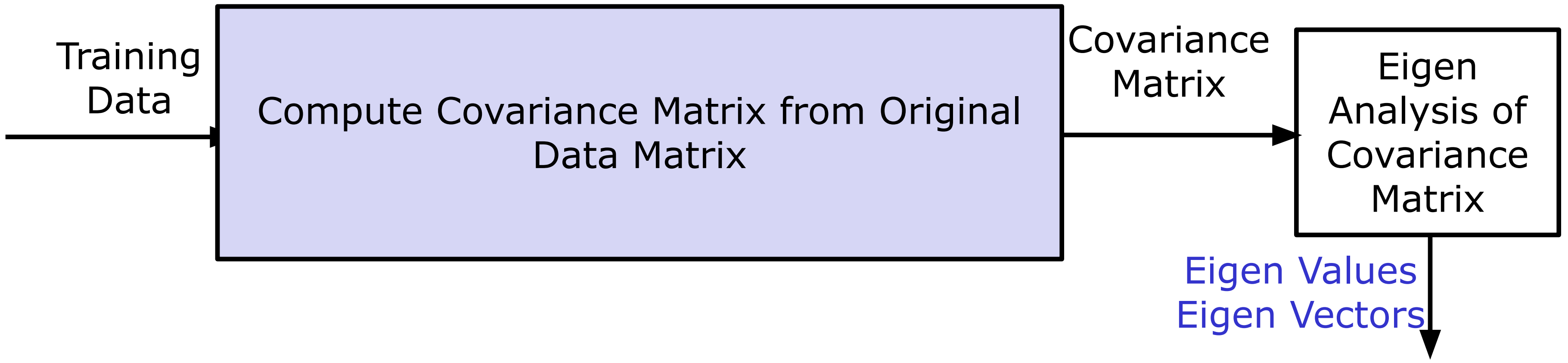
PCA during Model Building

- Model building and prediction using machine learning involve two stages:
  - Training stage: Model building
  - Test stage: Prediction using the built model
- Training stage: Perform the PCA on training data
  - Obtain the l direction of projection (eigenvectors) corresponding to l significant eigenvalues
  - Obtain the reduced dimension representation of training data by projecting training data on to l eigenvectors
- Test stage:
  - Obtain the reduced dimension representation of test data (test data vector(s)) by projecting it on to l eigenvectors obtained during training phase

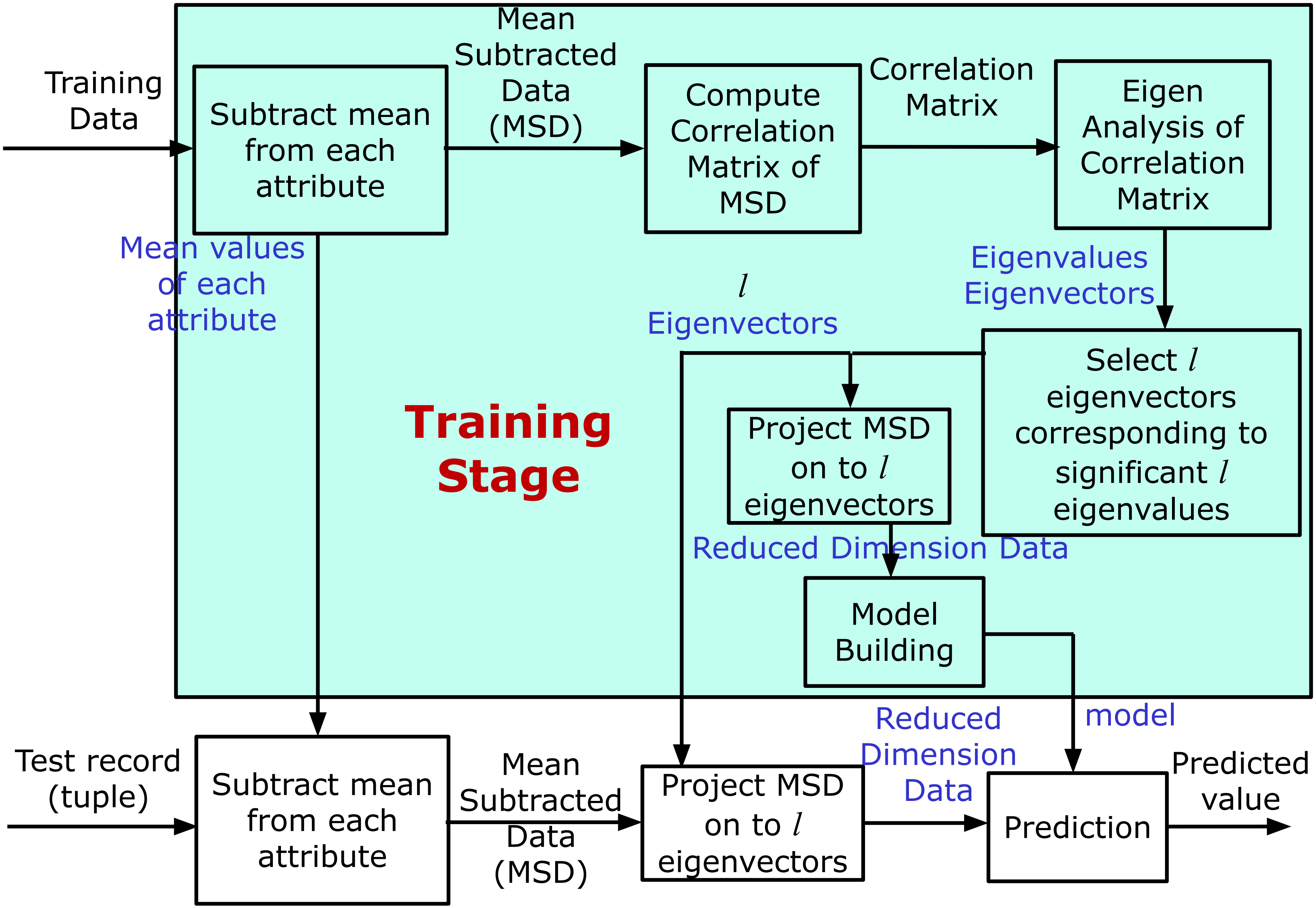
PCA during Model Building



PCA during Model Building



PCA during Model Building



Summary: Dimensionality Reduction

- This technique encodes (transforms) the original representation of data into a reduced or compressed representation of the original data
- Principal component analysis (PCA) is one of the popular and effective methods of lossy dimensionality reduction
- PCA can be used to obtain
  - uncorrelated reduced dimensional representation of data

    OR

  - uncorrelated transformed (encoded) representation with the number of dimension same as original data